

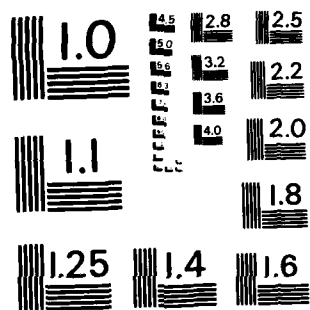
UNCLASSIFIED

BAYESIAN MODELS FOR RESPONSE SURFACES OF UNCERTAIN  
FUNCTIONAL FORM(U) WISCONSIN UNIV-MADISON MATHEMATICS  
RESEARCH CENTER D M STEINBERG JAN 83 MRC-TSR-2474  
DAAG29-80-C-0041 F/G 12/1

1/1

NL

END  
DATE  
FILMED  
6 83  
DTIC



MICROCOPY RESOLUTION TEST CHART  
NATIONAL BUREAU OF STANDARDS - 1963 - A

DA 127703

MRC Technical Summary Report # 2474

BAYESIAN MODELS FOR RESPONSE SURFACES  
OF UNCERTAIN FUNCTIONAL FORM

David M. Steinberg

Mathematics Research Center  
University of Wisconsin-Madison  
610 Walnut Street  
Madison, Wisconsin 53706

January 1983

(Received January 5, 1983)

DTIC FILE COPY

Approved for public release  
Distribution unlimited

Sponsored by

U. S. Army Research Office  
P. O. Box 12211  
Research Triangle Park  
North Carolina 27709

DTIC  
ELECTE  
MAY 06 1983  
S D E

83 05 06-143

UNIVERSITY OF WISCONSIN - MADISON  
MATHEMATICS RESEARCH CENTER



BAYESIAN MODELS FOR RESPONSE SURFACES OF UNCERTAIN FUNCTIONAL FORM

David M. Steinberg

Technical Summary Report #2474

January 1983

ABSTRACT

For	
Unannounced	<input checked="checked" type="checkbox"/>
Justification	<input type="checkbox"/>
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A	

Experimental response functions are often approximated by simple empirical functions such as polynomials. Several methods for modeling such responses which take into account this approximate nature are described and are shown to be essentially equivalent. The models all involve a Bayesian analysis which reflects prior experimental belief about the ability of the empirical approximation to represent the true response function. The models are also related to Kalman filters. Implications of the models for statistical inference are examined with particular attention to estimating the response function. Numerical examples help illustrate the models. A general predictive check is developed to examine the consistency of the model with the observed data.

AMS (MOS) Subject Classifications: 62F15, 62J05

Key Words: Response Surfaces; Bayesian Linear Model; Polynomial Regression; Spline Functions; Bias; Predictive Distribution; Lack of Fit Test; Kalman Filter.

Work Unit Number 4 - Statistics and Probability

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

## SIGNIFICANCE AND EXPLANATION

Scientists and engineers often wish to predict the value of a response variable as a function of one or more input variables. When the form of the response function is unknown or is very complicated, it is common to approximate the true response function by an empirical graduating function such as a polynomial whose coefficients are estimated from observed data.

Several statistical models have been proposed recently which reflect the approximate nature of empirical graduating functions. The models all follow a Bayesian approach which incorporates the experimenter's prior beliefs as to the likely (in)adequacy of the graduating function to represent the true response function. The resulting prediction equation has two components: an estimated graduating function and a second function which may be interpreted as the estimated bias induced by the particular choice of graduating function. The inclusion of the bias component typically allows the prediction equation to follow the observed data more closely than the graduating function alone. The extent to which the prediction equation will deviate from the graduating function depends largely on the experimenter's prior beliefs: if these express perfect confidence in the ability of the graduating function to model the response, then the bias will be estimated to be 0 and the prediction equation will contain only a graduating function; on the other hand, if the graduating function is thought to be seriously inadequate, the prediction equation will approximately interpolate the observed data points.

An example is given in which yield of a chemical process is thought to be roughly a linear function of the reaction temperature. The Bayesian methods produce prediction equations which show an overall linear increase of yield with temperature, but with local deviations about the trend line to more closely follow the observed experimental data.

---

The responsibility for the wording and views expressed in this descriptive summary lies with MRC, and not with the author of this report.

BAYESIAN MODELS FOR RESPONSE SURFACES  
OF UNCERTAIN FUNCTIONAL FORM

David M. Steinberg

1. Introduction

Many scientific investigations are designed to explore the relationship between a response variable  $Y$  and a set of input, or explanatory variables,  $\xi_1, \xi_2, \dots, \xi_k$ . The inputs may be continuous variables, such as dosage or time, or categorical variables, such as sex or batch number. Information on how the response is related to the inputs is obtained by conducting an experiment:  $n$  combinations of the inputs are specified and at each of these combinations an experiment is conducted and the resulting value of the response is observed.

Sometimes the physical nature of the problem suggests a specific functional form linking the response to the explanatory variables. However, in many investigations the functional nature of the response is either unknown, or is too complicated to provide a useful representation. A common strategy in such situations is to approximate the true, but unknown, response function by a simple graduating function, typically a low degree polynomial. For example, one might employ the first order approximation:

$$(1.1a) \quad Y(\xi) = \beta_0 + \sum \beta_i X_i,$$

where the  $q$  variables  $X_1, \dots, X_q$  are currently thought to be appropriate functions of the input variables  $\xi_1, \dots, \xi_k$  and  $\beta_0, \beta_1, \dots, \beta_q$  are parameters which must be estimated. This type of approximation is most common when the input variables are continuous but is also applicable for categorical inputs. If a first order approximation is judged to be an inadequate representation of the response function, one might use instead the second order approximation:

$$(1.1b) \quad Y(\xi) = \beta_0 + \sum \beta_i X_i + \sum \beta_{ij} X_i X_j.$$

When these equations are used to approximate the relationship between the response and the inputs, we can represent the experimental results by a linear statistical model:

$$(1.2) \quad Y = X\beta + \epsilon,$$

where  $Y$  denotes the  $n \times 1$  vector of observed responses,  $X$  is an  $n \times p$  matrix of regressors whose columns correspond to appropriate powers of the  $q$  functions of the input variables,  $X_1, \dots, X_q$ ,  $\beta$  is a vector of parameters which must be estimated, and  $\epsilon$  denotes the experimental error associated with the observations. The statistical techniques of response surface methodology can now be used to estimate the parameters and to analyze the behavior of the response (see Box and Wilson [1951], Box [1954], Box and Youle [1955], Myers [1976]).

Because (1.2) is only an approximate model based on an approximate form for the response function, it is natural to wonder what the effect might be of analyzing the experiment as though the approximating model were in fact the true response function. Box and Draper [1959] found that criteria for experimental design are affected dramatically when the approximate nature of the model is taken into account.

Recently, several authors have proposed generalizations of (1.2) which are designed to model scientific uncertainty about the form of the true response function. Smith [1973] suggests a three tiered Bayesian model following the general structure described by Lindley and Smith [1972]. Blight and Ott [1975] expand (1.2) by adding an extra term to reflect the difference between the approximation and the true model. O'Hagan [1978] proposes a "localized regression" model, in which the structure of (1.2) is maintained, but the parameter values are allowed to vary throughout the design space, subject to prior assumptions about their joint distribution. Wahba [1978]

advocates the use of approximating spline functions, and demonstrates how these generalize (1.2) and also how they can be viewed as Bayesian solutions to the estimation problem. Sacks and Ylvisaker [1978] suggest a model similar in form to that of Blight and Ott, but employ a frequentist, rather than a Bayesian, estimation procedure.

This paper will discuss the first four approaches, all of which involve Bayesian analyses. The next section will define the four models and will demonstrate how they are related to one another. Sections 3 and 4 will describe how the models can be used to draw inferences about the relationship between the response variable and the inputs. Section 5 will consider the traditional statistical test for lack of fit (i.e., model inadequacy) in the context of the Bayesian models. Section 6 will demonstrate some similarities and differences between these models and the Kalman filter. Some examples will be presented in Section 7 and concluding remarks are reserved for the final section.

## 2. Models for Unknown Response Functions

Smith [1973] proposes a hierarchical Bayesian linear model to represent the relationship between a response vector  $Y$  and a matrix  $X$  of associated regressor variables. The hierarchical structure consists of three tiers and provides the mechanism for building uncertainty about the response function into the statistical model. The model is a special case of the general three tiered Bayesian model analyzed by Lindley and Smith [1972] so that all of their results may be applied here.

The first tier of Smith's model simply states that the observation vector  $Y$  follows a multivariate normal distribution with mean vector  $\theta_1$  and variance matrix  $\sigma^2$  times the identity matrix, where  $\sigma^2$  indicates the magnitude of experimental error:

$$(2.1a) \quad Y/\theta_1 \sim N(\theta_1, \sigma^2 I).$$

The second tier invokes the linear model structure by asserting that the vector of expected values,  $\theta_1$ , follows a multivariate normal distribution with mean vector  $X\theta_2$ , where  $\theta_2$  is a vector of regression coefficients and corresponds directly to  $\beta$  in equation (1.2):

$$(2.1b) \quad \theta_1/\theta_2 \sim N(X\theta_2, V).$$

Smith examines in detail only the case where  $X\theta_2$  represents polynomial regression with a single regressor, but the generalization to arbitrary linear models is straightforward. The variance matrix  $V$  indicates the experimenter's a priori confidence in the adequacy of the linear model. Note that if the elements of  $V$  are all quite small, then the model is claiming that the expected value vector,  $\theta_1$ , follows the linear model  $X\theta_2$  very closely; i.e., the linear model is assumed to be a very good representation of the true response function. In fact, the limiting case of Smith's model as  $V$  converges to a 0 matrix yields precisely the same analysis that results from treating

(1.2) as an exact model for the experiment. On the other hand, if the elements of  $V$  are rather large, this reflects prior belief that the true response may deviate considerably from the linear model, even though it may be our best current guess for the response function. Just what constitutes "large" or "small" entries in  $V$  will be left vague for now; however, a reparameterization which makes these terms more precise will be described in section 3.

The final tier of the model assigns a prior distribution to the regression parameters:

$$(2.1c) \quad \theta_2 \sim N(\beta_0, V_1).$$

A diffuse prior is often deemed appropriate for the regression parameters and this can be achieved by considering limiting forms as  $V_1^{-1}$  converges to 0.

Blight and Ott [1975] propose a model which represents the response as a sum of three components:

$$(2.2) \quad \begin{aligned} \text{Response} = & \text{Low degree polynomial approximation} \\ & + \text{deterministic error (bias)} \\ & + \text{random (experimental) error.} \end{aligned}$$

The first and last terms are identical to the two terms in (1.2); what distinguishes this model from (1.2) is the second term, which is an explicit statement of the approximate nature of the polynomial. Actually, this type of statistical representation is well established: it is common practice to analyze estimators post-hoc by taking the expected value of the squared difference between the estimator and the quantity to be estimated and then to decompose this expected mean square error into variance and bias components. Blight and Ott merely suggest that bias, as well as variance, be incorporated into the original statistical model.

Mathematically, Blight and Ott's model can be written:

$$(2.3) \quad Y = X\beta + \eta + \epsilon,$$

The three terms on the right hand side of (2.3) correspond to the respective components of (2.2). Blight and Ott, like Smith, reserve close scrutiny for the special case where only one regressor variable is used in the polynomial approximation, but there is no difficulty in generalizing to arbitrary linear models. Observe that (2.3) is identical to (1.2) but for the addition of the middle term,  $\eta$ , and the assumption that this term permits an exact representation of  $Y$ , so that an equals sign is now justified.

Blight and Ott complete their model specification by making the following distributional assumptions:

$$(2.4a) \quad \beta \sim N(\beta_0, Q)$$

$$(2.4b) \quad \eta \sim N(0, L^{-1})$$

$$(2.4c) \quad \epsilon \sim N(0, \sigma^2 I).$$

In addition, it is assumed that  $\beta$ ,  $\eta$  and  $\epsilon$  are distributed independently.

A simple rationale underlies the distributional assumptions. Equation (2.4a) provides the prior distribution for the regression parameters and is directly analogous to (2.1c) in Smith's model. Again, a diffuse prior corresponds to considering limiting forms as  $Q^{-1}$  tends to a 0 matrix. The assumptions on  $\epsilon$  are identical to those in the standard linear model, that the experimental errors are independent and normally distributed with zero mean and common variance  $\sigma^2$ .

The distribution of the bias term given in (2.4c) is justified by appealing to prior belief. Recall that the bias term represents that part of the response function not captured by the approximating polynomial. Since the approximating polynomial typically represents the best current guess as to how the response depends on the inputs, it is reasonable to assign the bias a

prior mean of 0. The variance matrix of the bias suggests the possible severity of the bias; i.e., the prior adequacy assumed for the model. The diagonal elements of  $L^{-1}$  can be interpreted as reflecting the magnitude of the bias at the respective design points. The off-diagonal elements reflect prior assumptions about how similar the bias is likely to be at corresponding pairs of design points. This is closely related to prior convictions about the smoothness of the response function, since a response function which is smooth will have similar biases at proximate design points.

Both Smith and Blight and Ott propose to account for the approximate nature of the linear model (1.2) by adding an extra component. Smith suggests a hierarchical model, which assumes that the expected values of the observations need not exactly follow the linear model. Blight and Ott add a bias term which explicitly represents the difference between the true response function and the linear model. In both cases, a covariance matrix expresses the experimenter's prior belief about the degree to which the approximating model may be inadequate. It is clear that the two approaches share a similar spirit. We now show just how similar they are.

**Theorem 2.1** The Smith model and the Blight-Ott model are mathematically equivalent.

**Proof:** The proof is very simple and relies on a trivial re-writing of Smith's model. We will simply write each of the first two stages in Smith's model as the sum of a deterministic term (the expected value) plus a random term with an appropriate covariance matrix. (In fact, this is one of the methods employed by Lindley and Smith [1972] in deriving some of the properties of their more general hierarchical model.) Thus, we rewrite equation (2.1a) as:

$$(2.5a) \quad Y = \theta_1 + \epsilon, \quad \text{where } \epsilon \sim N(0, J^{-1}I).$$

Similarly, we rewrite equation (2.1b) as:

$$(2.5b) \quad \theta_1 = \mathbf{x}\theta_2 + \eta, \text{ where } \eta \sim N(0, \mathbf{V}).$$

Now, substituting (2.5b) into (2.5a) gives:

$$(2.5c) \quad \mathbf{Y} = \mathbf{x}\theta_2 + \eta + \epsilon,$$

where the distributions of  $\eta$  and  $\epsilon$  are given above and the distribution of  $\theta_2$  is given in (2.1c), and the three terms are independent. This is precisely the model suggested by Blight and Ott, with  $\theta_2$  in place of  $\beta$  and  $\mathbf{V}$  corresponding to  $\mathbf{L}^{-1}$ . This completes the proof.

The mathematical equivalence proved above is interesting in light of the somewhat different interpretations suggested for the two models. The model inadequacy approach advocated by Smith and the bias approach used by Blight and Ott would seem to be just two sides of the same coin.

O'Hagan [1978] suggests a different way to modify (1.2) to reflect uncertainty as to the form of the response function. He argues that, while (1.2) may be adequate to describe the response function in the immediate neighborhood of any particular point  $\mathbf{x} = (x_1(\xi), \dots, x_q(\xi))$ , it is unlikely to be valid over the entire range of inputs which might be used. This leads him to generalize (1.2) by allowing the parameter vector  $\beta$  to be a function of  $\mathbf{x}$ . The manner in which  $\beta$  varies with  $\mathbf{x}$  is not specified in terms of an exact functional form; rather, a prior probability distribution is used to describe the experimenter's beliefs as to how the parameters may change from one point to another. O'Hagan calls this the "localized regression model".

O'Hagan formally defines the localized regression model by specifying the appropriate distributional assumptions for each point  $\mathbf{x}$  in the design space. Denoting by  $Y_{\mathbf{x}}$  an observation at the point  $\mathbf{x}$ , he assumes that:

$$(2.6a) \quad Y_{\mathbf{x}}/\beta(\mathbf{x}) \sim N(f(\mathbf{x})'\beta(\mathbf{x}), \sigma^2),$$

where  $f(\mathbf{x})$  is a vector whose elements are the appropriate regressor variables

evaluated at  $x$  and  $\sigma^2$  is the experimental error variance. Further, he assumes that:

$$(2.6b) \quad \beta(x)/b_0 \sim N(b_0, B_0).$$

We can interpret  $b_0$  as the parameters of a global regression function about which there is local variation;  $B_0$  is a corresponding covariance matrix whose entries indicate how much the regression parameters at any particular point may vary about those for the global regression function. When there is no prior information to suggest a specific global regression function, O'Hagan suggests using a vague prior distribution for  $b_0$ . This is accomplished by assuming that:

$$(2.6c) \quad b_0 \sim N(b^*, kB^*),$$

and considering limiting forms as  $k \rightarrow \infty$ . Finally, he assumes that the parameter values for any two points  $x_1$  and  $x_2$  are stochastically related:

$$(2.6d) \quad E\{[\beta(x_1) - b_0][\beta(x_2) - b_0]'/b_0\} = V(x_1, x_2).$$

The matrix  $V$  in (2.6d) plays a similar role to the matrix  $L^{-1}$  in the Blight-Ott model. O'Hagan uses this matrix to reflect prior belief about how much the parameters (less the global parameter values) may vary from one point to another, while Blight and Ott wish to model how much the response function itself (less the polynomial approximation) may vary. In general, the entries of  $V$  are related to prior convictions about the smoothness of the response function: assuming that the parameter values at two design points are highly correlated is tantamount to assuming that a single approximating polynomial is adequate to describe the response function throughout the range between them. This idea can be used to suggest specific forms for  $V$ . For example, in the case of polynomial regression with a single regressor variable, O'Hagan suggests using:  $V(x_1, x_2) = \rho(|x_1 - x_2|)B_0$ , where  $\rho(d)$  is a monotone decreasing function of  $d$  and  $\rho(0)=1$ . For this choice,  $\beta(x)$  is a second-

order, stationary stochastic process on the real line and the rate at which  $\rho(d)$  decreases reflects prior assumptions about the smoothness of the response function. The slower the rate of decrease, the more highly correlated will be the parameter values at similar points, and hence the greater the degree of smoothness assumed for the response function.

From the above comments, it is clear that O'Hagan's model shares some common ground with Blight and Ott's model. In fact, O'Hagan observes that the Blight-Ott model "shows many similarities" to his own. However, concentrating on the specific covariance function analyzed in detail by Blight and Ott, he continues that "their model can be regarded as a special case of ours in that a localized regression term (locally constant) is added to a global term (polynomial). Their analysis relies on a special covariance structure for  $[\eta]$  and...their estimation of the global term takes no account of the average effect of the localized disturbance over the region of interest." By considering Blight and Ott's model in the more general form described in (2.3) and (2.4), we now show that it is actually equivalent to O'Hagan's model. Of course, by Theorem 2.1, this is also true of Smith's model.

**Theorem 2.2:** Under the model specification of (2.6a-d), the observed data vector  $\mathbf{Y}$  is described by the model (2.3)-(2.4), with  $b_0$  in place of  $\beta$  and

$$L_{ij}^{-1} = f(x_i)'V(x_i, x_j)f(x_j).$$

**Proof:** The proof is very similar to that of Theorem 2.1. We begin by rewriting (2.6a) as:

$$(2.7a) \quad Y_x = f(x)' \beta(x) + \epsilon_x, \quad \text{where } \epsilon_x \sim N(0, \sigma^2).$$

Now rewrite (2.6b) as:

$$(2.7b) \quad \beta(x) = b_0 + \zeta(x), \quad \text{where } \zeta(x) \sim N(0, B_0).$$

Substituting (2.7b) into (2.7a):

$$\begin{aligned} (2.7c) \quad Y_x &= f(x)'b_0 + f(x)'\zeta(x) + \epsilon_x \\ &= f(x)'b_0 + \eta_x + \epsilon_x, \text{ where } \eta_x = f(x)'\zeta(x). \end{aligned}$$

Thus the observation vector  $Y$  can be written:

$$(2.7d) \quad Y = Xb_0 + \eta + \epsilon,$$

precisely the form of (2.3). All that remains to complete the proof is to show that  $\eta$  has the distribution claimed in the theorem:

$$(2.7e) \quad E\{\eta\} = 0, \text{ because } E\{\zeta(x)\} = 0 \text{ for all } x.$$

$$\begin{aligned} (2.7f) \quad E\{[f(x_i)'\zeta(x_i)][f(x_j)'\zeta(x_j)]'\} \\ &= f(x_i)' E\{\zeta(x_i)\zeta(x_j)'\} f(x_j) \\ &= f(x_i)'\nabla(x_i, x_j)f(x_j). \end{aligned}$$

The last line follows from combining assumption (2.6d) of the model with the definition of  $\zeta(x)$  in (2.7b).

The spline function approach, on the surface, appears quite unrelated to the models described above. However, results of Wahba [1978] show that it is essentially an equivalent procedure when the regression coefficients are assigned a diffuse prior.

Generalized smoothing splines for the statistical problem posed here are derived as solutions to a problem in functional approximation. The solution in the general case exploits the structure of reproducing kernel Hilbert spaces (r.k.h.s.) (see Aronszajn [1950] for the general theory of r.k.h.s.). First, denote by  $\{\phi_j\}_{j=1}^P$  the monomials which constitute the approximating polynomial. Now let  $H_K$  be a r.k.h.s. of functions defined on the design space which contains the  $\{\phi_j\}_{j=1}^P$  and has reproducing kernel  $K(x_1, x_2)$ . It can be shown that  $H_K$  has a representation as the direct sum of span  $\{\phi_j\}$  and a r.k.h.s.  $H_Q$ , which has reproducing kernel  $Q(x_1, x_2)$ . Let  $P_Q$  be the orthogonal projection operator from  $H_K$  onto  $H_Q$ . Then the generalized smoothing spline

$g_{n,\lambda}$  is defined as the solution to the problem: find  $g \in H_K$  to minimize

$$(2.8) \quad n^{-1} \sum_{i=1}^n [g(x_i) - y_i]^2 + \lambda \|P_Q g\|_K^2$$

where the summation is over the  $n$  observed data points and the latter term is the squared norm (in  $H_K$ ) of the projection of  $g$  onto  $H_Q$  times a smoothing parameter  $\lambda$ .

Much of the work on smoothing splines has focused on the case where the design space is one dimensional,  $\phi_j = x^{j-1}$ ,  $j=1, \dots, p$ , and

$P_Q(g) = d^p g / dx^p$ . In this case, it is well known that  $g_{n,\lambda}$  is a polynomial spline of degree  $2p-1$  and is uniquely determined provided the data cannot be exactly interpolated by the approximating polynomial (see Wahba [1978]). A common choice for  $p$  has been  $p=2$ , in which case  $\|P_Q(g)\|^2 = \int (g^{(2)}(x))^2 dx$  and has a direct interpretation as a measure of the smoothness of the solution. The choice of  $\lambda$  then controls the tradeoff between how smooth the solution will be and how closely it will match the observed data.

The solution  $g_{n,\lambda}$  to (2.8) will always include the postulated approximating polynomial, since  $P_Q(\phi_j) = 0$ ,  $j=1, \dots, p$ . Thus, the polynomial can be included in the solution at no cost to the second term in (2.8), with the coefficients chosen to minimize the sum of squares. This prompts Wahba to suggest that "spline smoothing is an appropriate solution to the problem arising when one wants to fit a given set of regression functions to the data but one also wants to 'hedge' against model errors".

Wahba [1978] proves the following theorem which relates spline smoothing to Bayesian estimation of a stochastic process.

**Theorem 2.3:** Suppose the true response function is  $g(x)$ , so that the  $i$ 'th data point is

$$y_i = g(x_i) + \epsilon_i$$

where  $\epsilon = (\epsilon_1, \dots, \epsilon_n)' \sim N(0, \sigma^2 I)$ . Let the prior distribution of  $g(x)$  be

the same as that of the stochastic process

$$(2.9) \quad T_{\xi}(x) = \sum_{j=1}^p \beta_j \phi_j(x) + b^{1/2} Z(x),$$

where  $\beta = (\beta_1, \dots, \beta_p)' \sim N(\beta_0, \xi I)$ ,  $b > 0$  is fixed and  $Z(x)$  is a zero mean Gaussian stochastic process with  $E\{Z(x_1)Z(x_2)\} = Q(x_1, x_2)$ . Then for any fixed point  $x$ ,

$$g_{n,\lambda}(x) = \lim_{\xi \rightarrow \infty} E_{\xi}\{g(x)/Y=Y\},$$

where  $\lambda = \sigma^2/nb$  and  $E_{\xi}$  denotes expectation with respect to the posterior distribution of  $g(x)$  given the prior (2.9). Thus the smoothing spline solution  $g_{n,\lambda}$  is the limiting posterior expectation of the response function given (2.9) when the prior distribution of the parameters in the approximating polynomial is made diffuse.

The characterization of spline smoothing in Theorem 2.3 as a form of Bayesian estimation suggests a similarity with the previous models. We prove this in the following theorem.

**Theorem 2.4:** Under the prior specification (2.9) of the last theorem, the prior distribution of the data vector  $Y$  is given by the Blight-Ott model ((2.3) and (2.4)) with  $Q = \xi I$  and with  $L^{-1}_{i,j} = bQ(x_i, x_j)$ .

**Proof:** The  $i$ 'th observation is  $Y_i = g(x_i) + \epsilon_i$ . Then, given (2.9), the prior distribution for the  $i$ 'th observation is the same as the distribution of

$$\begin{aligned} \sum_{j=1}^p \beta_j \phi_j(x_i) + b^{1/2} Z(x_i) + \epsilon_i \\ = \sum_{j=1}^p \beta_j \phi_j(x_i) + \eta_i + \epsilon_i. \end{aligned}$$

The full data vector  $Y$  thus has a prior distribution identical to the distribution of

$$X\beta + \eta + \epsilon$$

where  $X$  is an  $n \times p$  matrix with  $X_{i,j} = \phi_j(x_i)$ ,  $\beta = (\beta_1, \dots, \beta_p)'$  and

$\eta = (\eta_1, \dots, \eta_n)'$ . The prior distributions of  $\beta$ ,  $\eta$ , and  $\epsilon$  are easily seen to be those claimed in the theorem.

The essence of Theorems 2.3 and 2.4 is that spline estimation can be derived as a Bayes estimate of the response function, and the prior specification which corresponds to spline estimation is essentially that proposed by Blight and Ott (or, equivalently, by Smith or O'Hagan). However, two qualifications are necessary. First, the Bayesian model defined by (2.9) prescribes the prior distribution of the response function for all possible factor settings. This is also true of O'Hagan's localized regression model. The Smith model of (2.1) and the Blight-Ott model of (2.3) and (2.4) are limited to the  $n$  observed data points. However, the natural extension of their models to draw inferences about the response at arbitrary factor combinations corresponds precisely to the priors of (2.6) and (2.9). Thus these "complete" priors, although not stated directly by Smith or by Blight and Ott, are nonetheless implicit in their models. This will be examined in detail in section 3.

The second difference between the model formulations concerns the matrix  $L^{-1}$ . In the Smith and Blight-Ott models, the entries of this matrix reflect the experimenter's prior belief as to the magnitude of bias at the design points and, in theory, they are restricted only by the requirement that  $L^{-1}$  be a legitimate covariance matrix. This is not so for O'Hagan's model or for spline estimation. The form of this matrix for O'Hagan's model was derived in Theorem 2.2 and was found to depend on the form of the approximating polynomial. The covariance matrix for spline estimation, given in Theorem 2.4, derives from the r.k.h.s. structure in conjunction with the choice of the approximating polynomial. The problem must be phrased in terms of an overall r.k.h.s.  $H_K$  which in turn must be decomposable into the direct sum of  $\text{span}\{\phi_j\}$  and a r.k.h.s.  $H_Q$ . The covariance matrix is then determined by the kernel  $Q(x_1, x_2)$  of  $H_Q$ . This restriction may result in some loss of

generality. In actual practice, the Bayesian interpretation of  $L^{-1}$  as a covariance matrix has played a secondary role in spline fitting. The primary concern has been to define an appropriate smoothness penalty. The smoothness penalty determines the appropriate r.k.h.s. and  $L^{-1}$  is then obtained as the Gramian matrix of appropriate elements of the r.k.h.s. (see Kimeldorf and Wahba [1971] for details). This is a valuable reminder that the basic motivation for using smoothing splines derives from ideas in functional approximation, not from its characterization as a Bayesian technique, although the latter has been emphasized here.

### 3. Inference for Future Observations: Prediction and Estimation

This section will discuss how statistical inferences may be drawn from the Bayesian models of section 2. The analysis will focus on the problem of predicting the response for various combinations of the input variables. This, after all, is the ultimate objective of the response surface framework described in section 1. However, we will also consider questions of inference regarding the regression coefficients themselves, in particular as they relate to predicting the response.

The notation used will be that of the Smith model (equation 2.1), although both the Smith model and the Blight and Ott models will prove useful in deriving some of the results. Of course, they will be equally valid for both models, as well as for O'Hagan's model and for the spline formulation when a diffuse prior is used for the regression coefficients. Most of the results will be stated as theorems. However, the proofs of the theorems will be deferred to the appendix at the conclusion of this report.

Consider first the problem of predicting the response  $Y$  when the input variables are fixed at  $\xi_1=c_1, \dots, \xi_k=c_k$ . First, let us convert the inputs to the more meaningful scale of  $X_1, \dots, X_q$ . We will then derive results for predicting  $Y(c)$  indirectly, in terms of  $Y(x)$ , where  $x=(X_1(c), \dots, X_q(c))$ .

For the Bayesian models of section 2, a natural method of prediction is to extend the model to include the prediction site, as well as the design points, and then to find the conditional expectation of  $Y$  at the prediction site given the observed data. Let  $Y$  denote the  $n$  observed experimental responses and let  $Y_x$  denote a (as yet unobserved) response at the prediction site. Similarly, denote by  $\theta_1$  and  $\theta_x$  the respective first tier expected values of  $Y$  and  $Y_x$ . Assume that the observational error associated with  $Y_x$ , like those in the experiment, is independent and normally distributed with

mean 0 and variance  $\sigma^2$ . Denote the regressor variables for  $\mathbf{Y}$  by an  $n \times p$  matrix  $\mathbf{X}$  and denote those for  $\mathbf{Y}_x$  by a  $1 \times p$  vector  $\mathbf{z}'$ . Denote the covariance matrix of the second tier by

$$\mathbf{V}^* = \begin{bmatrix} \mathbf{V} & \vdots & \mathbf{V} \\ \mathbf{V}' & \ddots & \mathbf{V}' \end{bmatrix},$$

where the partitioning corresponds to that of  $\theta_1$  and  $\theta_x$ . The entire model is now specified by:

$$(3.1a) \quad (\mathbf{Y}', \mathbf{Y}_x') / (\theta_1', \theta_x') \sim N((\theta_1', \theta_x'), \sigma^2 \mathbf{I}_{n+1})$$

$$(3.1b) \quad (\theta_1', \theta_x') / \theta_2 \sim N((\mathbf{X}\theta_2)', (\mathbf{z}'\theta_2)'), \mathbf{V}^*)$$

$$(3.1c) \quad \theta_2 \sim N(\beta_0, \mathbf{V}_1).$$

It should be noted that this is precisely the model that would be implied by the "more complete" prior specification of (2.9), when the prior for the regression coefficients,  $\theta_2$ , is made diffuse. This explains the comment of the previous section that the more complete prior is implicitly assumed in the Smith and Blight-Ott models.

**Theorem 3.1:** Under the model specification of (3.1), the conditional expectation of  $\mathbf{Y}_x$ , given that the observed data are  $\mathbf{Y}=\mathbf{y}$ , is:

$$(3.2) \quad E\{\mathbf{Y}_x / \mathbf{Y}=\mathbf{y}\} = \mathbf{z}'\beta_0 + (\mathbf{v}' + \mathbf{z}'\mathbf{V}_1\mathbf{X}')(\sigma^2 \mathbf{I}_n + \mathbf{V} + \mathbf{X}\mathbf{V}_1\mathbf{X}')^{-1}(\mathbf{y} - \mathbf{X}\beta_0).$$

The conditional expectation of the regression parameters is:

$$(3.3) \quad E\{\theta_2 / \mathbf{Y}=\mathbf{y}\} = \beta_0 + \mathbf{V}_1\mathbf{X}'(\sigma^2 \mathbf{I}_n + \mathbf{V} + \mathbf{X}\mathbf{V}_1\mathbf{X}')^{-1}(\mathbf{y} - \mathbf{X}\beta_0).$$

The results of Theorem 3.1, although straightforward, are not particularly revealing; equations (3.2) and (3.3) do not provide much intuition. The following corollaries and theorems will help to clarify this situation. We begin by observing how (3.2) and (3.3) are related.

**Corollary 3.1.1:** (i) The prediction for  $\mathbf{Y}_x$  has a natural decomposition as the sum of an approximating polynomial whose coefficients are estimated by  $E\{\theta_2 / \mathbf{Y}=\mathbf{y}\}$  and a second term, which Blight and Ott call the correction for bias, depending on  $\mathbf{v}$ :

$$(3.4) \quad E\{Y_x/Y-Y\} = x'E\{\theta_2/Y-Y\} + v'(\sigma^2 I_n + V + XV_1X')^{-1}(Y - X\beta_0) \\ = x'E\{\theta_2/Y-Y\} + E\{\eta_x/Y-Y\},$$

where  $\eta_x$  denotes the bias contribution at  $x$ , following the notation of (2.3).

(ii) The bias term in (3.4) can be simply expressed in terms of the residuals at the design points when the approximating polynomial alone is fit. Let us denote that vector of predictions by  $\hat{Y}_{AP}$ ; that is,  $\hat{Y}_{AP} = XE\{\theta_2/Y-Y\}$ . In a similar fashion, denote the corresponding residual vector by  $\hat{e}_{AP} = Y - \hat{Y}_{AP}$ . Then  $E\{\eta_x/Y-Y\} = v'(\sigma^2 I_n + V)^{-1}\hat{e}_{AP}$ .

The corollary provides valuable insight into the role that the matrix  $V^*$  plays in predicting  $Y_x$ . The elements of this matrix reflect the prior covariance assumed for the bias (or model inadequacy) at appropriate pairs of points in the design space. Let us denote the covariance function over all such pairs by  $v(x_1, x_2)$ . Then  $v' = (v(x, x_1), \dots, v(x, x_n))$ . Substituting this into the second term of (3.4) yields:

$$(3.5) \quad E\{Y_x/Y-Y\} = x'E\{\theta_2/Y-Y\} + \sum_{i=1}^n a_i v(x, x_i),$$

where the coefficients  $a_i$  are estimated from the data. The prediction equation, as a function of  $x$ , combines the approximating polynomial with a linear combination of  $n$  functions which are completely determined by the form of the covariance function and the choice of design points.

This might provide useful guidelines for choosing the covariance function, since some choices may lead to especially appealing prediction equations while others may have undesirable consequences. For example, Blight and Ott considered the special case of univariate polynomial regression and suggested using  $v(x, z) = \rho^2 \lambda^{|x-z|}$ ,  $0 < \lambda < 1$ , which is the covariance function for a first order autoregressive process. It can be seen from (3.5) that this results in a prediction equation whose derivative is discontinuous at each design point. If it is believed that the response function has a continuous

derivative, then this covariance function may be a poor representation of prior belief. Similarly, Smith's prior specification for univariate regression, whereby  $V = \tau^2 I$ , is called into question.  $V$  will be proportional to the identity, in general, only if  $v(x,z)$  decreases rapidly as  $x$  and  $z$  become distant from one another. But then the  $n$  functions  $v(x, x_i)$  in (3.5) will each have a rather sharp "spike" around  $x_i$ . The resulting prediction function will deviate from the approximating polynomial only in the vicinity of the design points, but these deviations may be quite sharp. This also would seem to be an unlikely summary of prior belief about the nature of the response function.

There are several interesting limiting cases of the above formulas. We have already observed that it is not uncommon to assign a diffuse prior to the regression coefficients. This special case is treated by Lindley and Smith [1972] and by Blight and Ott. It is also the situation in which the spline theory can be related to these models. As noted by Lindley and Smith, the prior is made diffuse by allowing  $V_1^{-1}$  to tend to 0. The following theorem applies the results of Theorem 3.1 to this special case.

Theorem 3.2: Suppose a diffuse prior is assumed for the regression coefficients. Then the predicted response at  $x$  is:

$$(3.6) \quad \lim_{V_1^{-1} \rightarrow 0} E\{Y_x / Y=y\} = x' [X'M^{-1}X]^{-1} X'M^{-1} y + v' \{M^{-1} - M^{-1}X[X'M^{-1}X]^{-1}X'M^{-1}\}y,$$

where  $M = (\sigma^2 I_n + V)$ . The estimated regression coefficients are:

$$(3.7) \quad \lim_{V_1^{-1} \rightarrow 0} E\{\theta_2 / Y=y\} = [X'M^{-1}X]^{-1} X'M^{-1} y.$$

Note that for fixed  $\theta_2$  the sampling distribution for  $Y$  under (3.1) is  $Y \sim N(X\theta_2, \sigma^2 I_n + V)$ . The maximum likelihood estimate of  $\theta_2$  under this model is precisely the estimate given in (3.7). Thus the well-known correspondence between maximum likelihood estimation of regression parameters and Bayesian

estimation with a diffuse prior is valid for these models.

As with Theorem 3.1, the above equations suggest a natural decomposition of the prediction equation into an approximating polynomial plus a "correction for bias" function:

Corollary 3.2.1: An alternative expression for the prediction equation when the regression parameters are assumed to have a diffuse prior is:

$$(3.8) \quad \lim_{\substack{\lambda \rightarrow 0 \\ \mathbf{V}^{-1} \rightarrow 0}} E\{Y_x / Y=y\} = \mathbf{z}' \lim_{\substack{\lambda \rightarrow 0 \\ \mathbf{V}^{-1} \rightarrow 0}} E\{\theta_2 / Y=y\} + \mathbf{v}' \mathbf{M}^{-1} \{ \mathbf{I} - \mathbf{X}(\mathbf{X}' \mathbf{M}^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{M}^{-1} \} \mathbf{y}.$$

Theorem 3.2 also suggests a different way to write (3.6) and (3.7), which directly contrasts the importance of bias and experimental error.

Corollary 3.2.2: Suppose that  $\mathbf{V}^*$  is a scaled version of a "standardized" covariance matrix  $\mathbf{R}^*$ ; that is,  $\mathbf{V}^* = \tau \mathbf{R}^*$ . Denote by  $\mathbf{R}$ ,  $\mathbf{r}$  and  $r$  the partitioned pieces of  $\mathbf{R}^*$  which correspond to  $\mathbf{V}$ ,  $\mathbf{v}$  and  $v$ , respectively, in  $\mathbf{V}^*$ . Let  $\lambda = \sigma^2 / \tau$ . Then:

$$(3.9) \quad \lim_{\substack{\lambda \rightarrow 0 \\ \mathbf{V}^{-1} \rightarrow 0}} E\{Y_x / Y=y\} = \mathbf{z}' (\mathbf{X}' \mathbf{C}^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{C}^{-1} \mathbf{y} + \mathbf{r}' \{ \mathbf{C}^{-1} - \mathbf{C}^{-1} \mathbf{X} (\mathbf{X}' \mathbf{C}^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{C}^{-1} \} \mathbf{y},$$

where  $\mathbf{C} = \lambda \mathbf{I}_n + \mathbf{R}$ , and

$$(3.10) \quad \lim_{\substack{\lambda \rightarrow 0 \\ \mathbf{V}^{-1} \rightarrow 0}} E\{\theta_2 / Y=y\} = (\mathbf{X}' \mathbf{C}^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{C}^{-1} \mathbf{y}.$$

Corollary 3.2.2 is important in that it shows the contrasting roles of experimental error and bias in determining the prediction equation (3.9) and the estimated regression coefficients (3.10): they depend on  $\sigma^2$  and  $\tau$  only through their ratio  $\lambda$ . This ratio thus provides a concise summary of how "large" or "small" are the elements of  $\mathbf{V}$ . Allowing  $\lambda$  to range from 0 to infinity permits us to model situations from those in which the bias is totally dominant (as might be the case in numerical analysis) through those in which the experimental error is dominant (when, say, scientific knowledge provides an exact form for the response function). This parameter corresponds directly to the smoothing parameter in the spline formulation (2.8) and we

shall see in Theorems 3.3 and 3.5 how it acts as a smoothing parameter here. (An essential assumption in obtaining this result is that the regression coefficients have a diffuse prior; if a proper prior is used, the ratio of  $\sigma^2$  to  $\tau$  does not determine the conditional expectations.)

An important conclusion of the corollary is that the degree to which a polynomial approximation  $P(\mathbf{x})$  is an adequate representation of a response function  $g(\mathbf{x})$  depends on the extent of experimental error, not just the absolute bias,  $g(\mathbf{x}) - P(\mathbf{x})$ . Thus, for example, even if the response function differs considerably from the form of the approximating polynomial, substantial experimental error will most likely make it impossible to detect this and our estimate of the response function should not deviate greatly from the approximating polynomial. On the other hand, if experimental error is very small, even minor departures from the approximating polynomial may be graphically obvious from a simple plot of the data and our estimator should be modified accordingly. This fundamental observation that the magnitude of bias must necessarily be evaluated relative to experimental error is also a basic feature of the design criteria developed by Box and Draper [1959].

The corollary suggests that  $\mathbf{V}^*$  be written as proportional to a "standardized" covariance matrix. Although there is no precise way of specifying what is meant by "standardized", we can offer some guidelines. One natural possibility, if the diagonal elements of  $\mathbf{V}^*$  are all assumed to be equal (as, for example, in Smith and in Blight and Ott), is to let  $\mathbf{R}^*$  be the correlation matrix which corresponds to  $\mathbf{V}^*$ ; both of the preceding papers used precisely such a parameterization. When the bias variance is thought to depend on  $\mathbf{x}$ , which might be appropriate for many response surface situations, then  $\mathbf{R}^*$  might be defined so that the variance for some specified  $\mathbf{x}$  would be equal to 1. Then  $\lambda$  would represent the ratio of experimental error to bias

at that particular  $\mathbf{x}$ .

Theorem 3.3: Suppose we can write  $\mathbf{V}^* = \tau \mathbf{R}^*$ , as in Corollary 3.2.2. Then we have the following limiting forms as  $\lambda \rightarrow \infty$ :

$$(3.11) \quad \lim_{\lambda \rightarrow \infty} \lim_{\mathbf{V}_1^{-1} \rightarrow 0} E\{Y_{\mathbf{x}} / \mathbf{Y} = \mathbf{y}\} = \mathbf{z}' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y}.$$

$$(3.12) \quad \lim_{\lambda \rightarrow \infty} \lim_{\mathbf{V}_1^{-1} \rightarrow 0} E\{\theta_2 / \mathbf{Y} = \mathbf{y}\} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y}.$$

If  $\mathbf{R}$  is non-singular, then we have the following limits as  $\lambda \rightarrow 0$ :

$$(3.13) \quad \lim_{\lambda \rightarrow 0} \lim_{\mathbf{V}_1^{-1} \rightarrow 0} E\{Y_{\mathbf{x}} / \mathbf{Y} = \mathbf{y}\} = \mathbf{z}' (\mathbf{X}' \mathbf{R}^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{R}^{-1} \mathbf{y} \\ + \mathbf{r}' \{ \mathbf{R}^{-1} - \mathbf{R}^{-1} \mathbf{X} (\mathbf{X}' \mathbf{R}^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{R}^{-1} \} \mathbf{y}, \text{ and}$$

$$(3.14) \quad \lim_{\lambda \rightarrow 0} \lim_{\mathbf{V}_1^{-1} \rightarrow 0} E\{\theta_2 / \mathbf{Y} = \mathbf{y}\} = (\mathbf{X}' \mathbf{R}^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{R}^{-1} \mathbf{y}.$$

The first half of Theorem 3.3 yields familiar answers: the ordinary least squares estimators. Thus, ordinary least squares obtains as a limiting case of the Bayesian model when the regression parameters have a diffuse prior and when the bias is assumed to be negligible relative to experimental error (i.e. when the approximating polynomial is assumed to exactly represent the response function). The second half of Theorem 3.3 does not have so immediate an interpretation. However, the following corollary makes clear what happens when  $\lambda$  tends to 0.

Corollary 3.3.1: Denote by  $\hat{\mathbf{Y}}$  the  $n \times 1$  vector of predicted values corresponding to the  $n$  design points; that is,  $\hat{Y}_i = E\{Y_{\mathbf{x}_i} / \mathbf{Y} = \mathbf{y}\}$ . If  $\mathbf{R}$  is a non-singular matrix, then:

$$\lim_{\lambda \rightarrow 0} \lim_{\mathbf{V}_1^{-1} \rightarrow 0} \hat{\mathbf{Y}} = \mathbf{y};$$

i.e. the prediction equation interpolates the observed data. If there are replicate observations at some of the design points, these will contribute

identical rows to  $R$ , making it singular. Although the corollary cannot then be applied directly, it is easy to use the corollary to show that the limiting prediction curve as  $\lambda \rightarrow 0$  interpolates the averages of the responses observed at each design point.

If we think of the prediction equation  $E\{Y_x/Y=y\}$  as a function of  $\lambda$ , we see that it varies from an interpolant (when  $\lambda=0$ ) to the least squares approximating polynomial (as  $\lambda$  tends to infinity). The former result seems intuitively reasonable, for  $\lambda=0$  describes the situation in which there is no experimental error, so that the observed responses are exact values of the response function. The prediction equation reflects this certain knowledge by correctly predicting the response at those points.

This particular characterization of the prediction equation as a function of  $\lambda$  is well-known in the spline literature (see, for example, Kimeldorf and Wahba [1971]). Blight and Ott were evidently unaware that it held for their model as well. They proposed a parametric form for the  $R$  matrix, and then suggested that these parameters and  $\lambda$  be jointly estimated by minimizing the residual sum of squares,  $S(R, \lambda) = \sum (y_i - \hat{y}_i)^2$ . It is clear from the corollary that this will always be minimized when  $\lambda = 0$ , regardless of the values of the other parameters.

We can also state some additional properties of the predicted value vector  $\hat{Y}$ . A common characterization of Bayes estimates is that they can be expressed as weighted averages of their prior means and the observed data. This is not possible for predicting the response at an arbitrary point  $x$  since in general no observation has been made there; however, for the  $n$  observed data points and for the estimated regression coefficients, we can write such weighted averages.

Theorem 3.4: The predicted value vector  $\hat{\mathbf{Y}}$  can be expressed as a weighted average of its prior mean,  $\mathbf{X}\beta_0$  and the observed responses,  $\mathbf{y}$ , where the weights are inversely proportional to the respective measures of variation,  $\mathbf{V} + \mathbf{XV}_1\mathbf{X}'$  and  $\sigma^2$ :

$$(3.15) \quad \hat{\mathbf{Y}} = [\sigma^2\mathbf{I}_n + (\mathbf{V} + \mathbf{XV}_1\mathbf{X}')^{-1}]^{-1} [\sigma^2\mathbf{y} + (\mathbf{V} + \mathbf{XV}_1\mathbf{X}')^{-1}\mathbf{X}\beta_0].$$

Similarly, the estimated regression coefficients can be written as a weighted average of their prior mean  $\beta_0$  and the observed responses, with the weights inversely proportional to the prior and data precision matrices, respectively:

$$(3.16) \quad E\{\theta_2/\mathbf{Y}=\mathbf{y}\} = [\mathbf{X}'(\sigma^2\mathbf{I}_n + \mathbf{V})^{-1}\mathbf{X} + \mathbf{V}_1^{-1}]^{-1} [\mathbf{X}'(\sigma^2\mathbf{I}_n + \mathbf{V})^{-1}\mathbf{y} + \mathbf{V}_1^{-1}\beta_0].$$

The following theorem demonstrates more clearly the link between the Bayesian models and the statistical smoothing approach.

Theorem 3.5:  $\hat{\mathbf{Y}}$  solves the minimization problem: find  $\mathbf{u}$  to minimize

$$(3.17) \quad (\mathbf{u}-\mathbf{y})'(\mathbf{u}-\mathbf{y}) + (\mathbf{u}-\mathbf{X}\beta_0)'[\sigma^2(\mathbf{V} + \mathbf{XV}_1\mathbf{X}')^{-1}](\mathbf{u}-\mathbf{X}\beta_0).$$

If  $\mathbf{V}_1^{-1} = 0$ , (3.17) becomes:

$$(3.18) \quad (\mathbf{u}-\mathbf{y})'(\mathbf{u}-\mathbf{y}) + \lambda\mathbf{u}'[\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}^{-1}]\mathbf{u},$$

where  $\lambda$  and  $\mathbf{R}$  are defined as in Corollary 3.2.2. Moreover, the second term in (3.18) is 0 if and only if  $\mathbf{u} \in \text{col}(\mathbf{X})$ , that is, if and only if  $\mathbf{u}$  can be written as a linear combination of the columns of  $\mathbf{X}$ .

Both (3.17) and (3.18) characterize the prediction vector  $\hat{\mathbf{Y}}$  as the solution to a minimization problem composed of two terms: the residual sum of squares,  $(\hat{\mathbf{Y}}-\mathbf{y})'(\hat{\mathbf{Y}}-\mathbf{y})$ , and a quadratic penalty term. For the general case, (3.17), the quadratic penalty is in terms of the distance of  $\hat{\mathbf{Y}}$  from its prior expectation,  $\mathbf{X}\beta_0$ . This will clearly have the effect of "shrinking" the vector of predicted values toward the prior expectation. The extent of this shrinkage depends on the weighting matrix  $\sigma^2(\mathbf{V} + \mathbf{XV}_1\mathbf{X}')^{-1}$ . This matrix is proportional to  $\sigma^2$ , but inversely proportional to the prior variance. Thus the prior expectation will be most important when our prior precision is great

relative to experimental error; when our prior precision is not great, the data will dominate the prior in determining  $\hat{\mathbf{Y}}$ .

The quadratic penalty term undergoes several interesting changes in the limiting case of (3.18). First, the penalty can be expressed solely in terms of the "standardized" bias covariance matrix  $\mathbf{R}$  and the variance-bias tradeoff parameter  $\lambda$ . Second, the penalty is independent of the prior expectation. Third, the penalty is 0 only for those vectors of predicted values which are in the column space of  $\mathbf{X}$ ; i.e. for those vectors of predicted values which can be exactly written as an approximating polynomial. The meaning of these last two points is that the penalty does not induce shrinkage toward a particular pre-specified vector; rather, in a more general sense, there is shrinkage toward the response plane spanned by the approximating polynomial. Finally, note that equation (3.18) is an exact discrete analogue of (2.8), the continuous smoothing problem which leads to generalized spline estimation. Were we to discretize the more general penalty function of (2.8) and limit it to the design points, we would obtain an equation like (3.18); unfortunately, it is impossible to follow this path in the reverse direction and derive a continuous penalty which corresponds to a discrete one. Nonetheless, (3.18) illustrates the close link between spline estimation and the Bayesian models with a diffuse prior on the regression coefficients.

When  $\mathbf{V}_1^{-1} = \mathbf{0}$ , Theorem 3.5 allows us to show how the residual sum of squares depends on the choice of  $\lambda$ . Let us denote the vector of predicted values by  $\hat{\mathbf{Y}}(\lambda)$  to emphasize its dependence on  $\lambda$ . Then define the residual sum of squares function by  $\text{RSS}(\lambda) = [\hat{\mathbf{Y}}(\lambda) - \mathbf{y}]' [\hat{\mathbf{Y}}(\lambda) - \mathbf{y}]$ .

Corollary 3.5.1:  $\text{RSS}(\lambda)$  is a monotone increasing function of  $\lambda$ , with

$\text{RSS}(\lambda=\infty)$  equal to the residual sum of squares from fitting the approximating polynomial by ordinary least squares and, provided  $\mathbf{R}$  is non-singular, with

$RSS(0) = 0.$

Additional comments on the type of estimates produced by these Bayesian models will be found in section 7, where several examples are presented.

#### 4. Inference for Future Observations: Precision

Statisticians are rarely satisfied with point estimates alone; some measure of the precision of the estimates is also essential. Since the Bayesian models base their estimates on posterior distributions of the quantities of interest, rather than on sampling properties, it is natural to obtain measures of precision from the posterior distributions, as well. Since the scope of this paper is limited to multivariate normal distributions, the natural measures of precision are the appropriate posterior variance matrices. These will be presented in this section.

Theorem 4.1: Let  $Y_{\mathbf{x}}$  denote a potential future observation at  $\mathbf{x}$ . Suppose the joint distribution of  $Y_{\mathbf{x}}$  and the observed response vector  $\mathbf{Y}$  is given by

(3.1). Then the posterior variance of  $Y_{\mathbf{x}}$  is:

$$(4.1) \quad \text{Var}\{Y_{\mathbf{x}}/\mathbf{Y}=\mathbf{y}\} = \sigma^2 + \mathbf{v} + \mathbf{z}'[\mathbf{X}'(\sigma^2\mathbf{I}_n + \mathbf{V})^{-1}\mathbf{X} + \mathbf{V}_1^{-1}]^{-1}\mathbf{z} \\ - 2\mathbf{v}'(\sigma^2\mathbf{I}_n + \mathbf{V} + \mathbf{XV}_1\mathbf{X}')^{-1}\mathbf{XV}_1\mathbf{z} \\ - \mathbf{v}'(\sigma^2\mathbf{I}_n + \mathbf{V} + \mathbf{XV}_1\mathbf{X}')^{-1}\mathbf{v}.$$

The posterior variance matrix for the regression coefficients is:

$$(4.2) \quad \text{Var}\{\theta_2/\mathbf{Y}=\mathbf{y}\} = [\mathbf{X}'(\sigma^2\mathbf{I}_n + \mathbf{V})^{-1}\mathbf{X} + \mathbf{V}_1^{-1}]^{-1}.$$

The posterior variance of  $Y_{\mathbf{x}}$  is composed of several components. One of these is, of course,  $\sigma^2$ , since any future observation is itself subject to observational error. Another component (the last term on the first line of (4.1)) is clearly seen to be  $\text{Var}\{\mathbf{z}'\theta_2/\mathbf{Y}=\mathbf{y}\}$ , the posterior variance of the approximating polynomial part of  $Y_{\mathbf{x}}$ . The remaining components involve quadratic forms of the bias covariances and the regressor variables, but do not suggest any obvious interpretation. Both of the posterior variances given above depend on the choice of the experimental design (i.e. on  $\mathbf{X}$ ), but not on the observed responses  $\mathbf{y}$ , a well-known property of conditional variances for multivariate normal distributions.

As in the last section, it is of particular interest to study the special case when the regression coefficients are assigned a diffuse prior.

Corollary 4.1.1: Given the assumptions of Theorem 4.1:

$$(4.3) \quad \lim_{\substack{\lambda \rightarrow 0 \\ V_1 \rightarrow 0}} \text{Var}\{Y_x/Y-y\} = \sigma^2 + v + z'(X'M^{-1}X)^{-1}z \\ - 2v'M^{-1}X(X'M^{-1}X)^{-1}z \\ - v'[M^{-1} - M^{-1}X(X'M^{-1}X)^{-1}X'M^{-1}]v,$$

where  $M = \sigma^2 I_n + V$ . Similarly:

$$(4.4) \quad \lim_{\substack{\lambda \rightarrow 0 \\ V_1 \rightarrow 0}} \text{Var}\{\theta_2/Y-y\} = (X'M^{-1}X)^{-1}.$$

Once again the use of a diffuse prior leads to the same variance matrix for the regression coefficients that would be obtained by traditional sampling theory methods for this model.

The formulas of Corollary 4.1.1 can be rewritten in a standardized form in precisely the same manner as the estimates of the preceding section. Following the same notation, let  $R$  denote the standardized bias covariance matrix, with  $V = \tau R$ , and let  $\lambda = \sigma^2/\tau$  be the variance-bias tradeoff parameter. Unlike the posterior expectation of  $Y_x$ , which depended on  $\sigma^2$  and  $\tau$  only through  $\lambda$ , the posterior variance proves to be proportional to  $\sigma^2$ , with the constant of proportionality a function of only  $\lambda$ .

Corollary 4.1.2: Under the assumptions of Theorem 4.1:

$$(4.5) \quad \lim_{\substack{\lambda \rightarrow 0 \\ V_1 \rightarrow 0}} \text{Var}\{Y_x/Y-y\} = \sigma^2 \{1 + \lambda^{-1}r + \lambda^{-1}z'(X'C^{-1}X)^{-1}z \\ - 2\lambda^{-1}r'C^{-1}X(X'C^{-1}X)^{-1}z \\ - \lambda^{-1}r'[C^{-1} - C^{-1}X(X'C^{-1}X)^{-1}X'C^{-1}]r\},$$

where  $C = \lambda I_n + R$ , and:

$$(4.6) \quad \lim_{\substack{\lambda \rightarrow 0 \\ V_1 \rightarrow 0}} \text{Var}\{\theta_2/Y-y\} = \sigma^2 \lambda^{-1} (X'C^{-1}X)^{-1}.$$

## 5. Testing For Lack Of Fit

The Bayesian models studied in this paper have been designed to describe situations in which an unknown response function is approximated by a simple graduating function such as a polynomial. The models seek to account for this approximate nature, whereas traditional analysis typically proceeds as though the graduating function were an exact representation of the response function. However, traditional analysis is not blind to the fact that a particular graduating function may be inadequate to represent a complex response function, and numerous diagnostic procedures have been developed to help us judge model inadequacy. One such procedure is the lack of fit test. This section will show how the lack of fit test relates to the Bayesian models of section 2.

The classical lack of fit test depends on the availability of an "independent" estimate of the error variance  $\sigma^2$ . (By independent, we mean an estimate which is independent of the assumed approximating function.) Such an estimate might result from knowledge of the experimental situation, past experience or, as is often the case, from the inclusion of replicate observations in the experiment. The extent to which the observed responses vary about the estimated graduating function can then be compared to this standard; if the variation is excessive, this is an indication that the approximating function is inadequate.

Specifically, suppose the approximating model (1.2) is exact, so that  $Y = X\beta + \epsilon$ , and suppose further that  $\epsilon \sim N(0, \sigma^2 I)$ . Let  $\hat{Y}$  denote the vector of predicted responses obtained from ordinary least squares regression. Then it is well known that the scaled residual sum of squares is distributed as a chi-squared random variable:

$$(5.1) \quad RSS/\sigma^2 \sim \chi_{n-p}^2,$$

where  $RSS = (\hat{Y} - Y)'(\hat{Y} - Y)$  is the residual sum of squares,  $n$  is the number of observations and  $p$  is the rank of  $X$ . The lack of fit test thus compares the observed RSS, scaled by the prior estimate of  $\sigma^2$ , with percentage points of the corresponding chi-squared distribution. When the estimate of  $\sigma^2$  is itself obtained from the experiment via replicate observations, the analogous procedure is to divide the RSS into two components:

$$(5.2) \quad RSS = SS(\text{Pure Error}) + SS(\text{Lack of Fit}),$$

where  $SS(\text{Pure Error})$  is the replication sum of squares and  $SS(\text{Lack of Fit})$  is the sum of squared deviations of the replicate averages about the estimated graduating function. These terms are divided by their appropriate degrees of freedom, and their ratio then has an F distribution under the assumption that the approximating model is correct.

The Bayesian models described in section 2 account for the approximate nature of the graduating function by adding an extra term. In section 3, we saw that when these models assign a diffuse prior to the regression coefficients, the extent to which the added term influences the estimates is controlled by the single parameter  $\lambda$ , the ratio of experimental error to bias. In particular, the limiting case of  $\lambda = \infty$  reduces to the traditional model in which the graduating function is assumed to exactly represent the response. Thus it seems reasonable that we should be able to generalize the lack of fit test to the Bayesian models as a test regarding  $\lambda$  in such a way that the test described above results in the limit as  $\lambda$  tends to  $\infty$ .

A general method for deriving diagnostic checks of parameters in Bayesian models has been suggested by Box [1980]. He advocates examining the consistency of the observed data vector  $Y$  with the predictive distribution implied for  $Y$  by the model. Of course, the predictive distribution of  $Y$  is simply the marginal distribution given in Lemma 3.1 of the appendix:

$$(5.3a) \quad Y \sim N(X\beta_0, \sigma^2 I + V + XV_1X').$$

Using the reparameterization suggested by the standardized form of Corollary 3.2.2, let us replace  $V$  by  $\lambda^{-1}\sigma^2 R$  to obtain:

$$(5.3b) \quad Y \sim N(X\beta_0, \sigma^2 I + \lambda^{-1}\sigma^2 R + XV_1X').$$

A logical diagnostic checking function based on the predictive distribution is the sum of squared deviations from the mean vector weighted by the inverse of the covariance matrix:

$$(5.4) \quad h(\lambda) = (Y - X\beta_0)'(\sigma^2 I + \lambda^{-1}\sigma^2 R + XV_1X')^{-1}(Y - X\beta_0).$$

If the model is true, then  $h(\lambda)$  should be distributed as a chi-squared random variable with  $n$  degrees of freedom. Assuming the other parameters in the model are exactly specified (including  $\sigma^2$ ), this can be used to check whether an hypothesized value for  $\lambda$  is consistent with the data.

We have already remarked that considerable interest focuses on the special case when the regression parameters are assigned a diffuse prior. We now examine how this affects the diagnostic checking procedure described above. To do so, we must first establish some simple results about the residual vector which results when a diffuse prior is used. Proofs for all of the following propositions are given in the appendix.

**Lemma 5.1:** Suppose  $Y$  follows the model described by (2.1), with  $V = \lambda^{-1}\sigma^2 R$ .

Let  $\hat{Y}(\lambda)$  denote the vector of predicted values when the regression coefficients are assigned a diffuse prior, and let  $\hat{e}(\lambda)$  denote the corresponding residual vector; i.e.  $\hat{e}(\lambda) = Y - \hat{Y}(\lambda)$ . Then:

$$(5.5) \quad \hat{e}(\lambda) = B(\lambda)Y,$$

where  $B(\lambda) = \lambda[C^{-1} - C^{-1}X(X'C^{-1}X)^{-1}X'C^{-1}]$  and  $C = \lambda I_n + R$ .

**Theorem 5.1:** Let us denote by  $h^*(\lambda)$  the special case of  $h(\lambda)$  which obtains when the regression parameters have a diffuse prior; i.e.

$$h^*(\lambda) = \lim_{V_1 \rightarrow 0} h(\lambda). \quad \text{Then:}$$

$$(5.6) \quad h^*(\lambda) = [B(\lambda)Y]'(I_n + \lambda^{-1}R)[B(\lambda)Y]/\sigma^2.$$

In the limit, as  $\lambda \rightarrow \infty$ , we have:

$$(5.7) \quad \lim_{\lambda \rightarrow \infty} h^*(\lambda) = \text{RSS}/\sigma^2,$$

where RSS denotes the residual sum of squares which results from ordinary least squares regression.

The limiting case given in (5.7) above is precisely the traditional lack of fit statistic suggested in (5.1) for situations in which  $\sigma^2$  is assumed to be known. Thus, the predictive check described above does provide a generalization of the traditional lack of fit test when the regression coefficients are assigned a diffuse prior. For general  $\lambda$  the diagnostic statistic given in (5.6) is a weighted, scaled, residual sum of squares, where the weighting matrix is  $I_n + \lambda^{-1}R = \lambda^{-1}C$  and the scale factor is  $\sigma^2$ .

An interesting question arises concerning the appropriate reference distribution with which to compare the diagnostic statistic (5.6). We obtained (5.6) as a limiting case of (5.4), which we argued earlier should have a chi-squared distribution with  $n$  degrees of freedom. However, when we examined the limiting case of (5.6) which corresponds to perfect faith in the approximating model, we obtained (5.7), the traditional lack of fit statistic, which has a sampling distribution which is chi-squared with  $n-p$  degrees of freedom, where  $p$  is the rank of  $X$ . What happened to the  $p$  degrees of freedom, and which, if either, of these distributions is an appropriate reference distribution for (5.6)?

It seems clear that of the two limiting processes which lead from (5.4) to (5.7), the more important is the use of the improper prior for the regression coefficients. This affects  $p$  dimensions of the prior distribution, precisely the number of degrees of freedom lost. We might then think of the

improper prior as forcing us to replace those  $p$  dimensions with information from the data, at the expense of information about the residuals. The results of Theorem 3.5 are also germane here. Theorem 3.5 characterized  $\hat{\mathbf{Y}}$  as a shrinkage type estimate. When the regression coefficients were assigned a proper prior, the shrinkage was in the direction of the prior mean,  $\mathbf{X}\beta_0$ ; however, when an improper prior was used, the shrinkage was in the direction of the  $p$ -dimensional regression plane, rather than toward a specific vector in that plane. Again, this suggests that using a data estimated  $\hat{\mathbf{Y}}$  rather than a prior mean vector reduces the dimension of the space in which the residuals lie, resulting in a loss of  $p$  degrees of freedom. Thus, although it is not formally true that  $h^*(\lambda)$  has a probability distribution (an inevitable consequence of using a vague prior for the regression coefficients), we feel that  $\chi^2_{n-p}$  is an appropriate reference distribution for the diagnostic statistic (5.6) which can be used to check any hypothesized value of  $\lambda$ .

**Lemma 5.2:** Consider the weighted residual sum of squares which appears in (5.6), but with the vector of predicted values, and hence the residual vector, determined by a proper prior. For any  $\lambda$ :

$$\lim_{\lambda \rightarrow 0} E\{(\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{I}_n + \lambda^{-1}\mathbf{R})(\mathbf{Y} - \hat{\mathbf{Y}})\} = (n-p)\sigma^2.$$

The above expression is not the expected value of  $h^*(\lambda)$  since we have reversed the limit and the expectation. (As noted above, formally,  $h^*(\lambda)$  does not have a probability distribution, and hence has no expectation.) Nonetheless, the lemma does provide additional support for the claim that the proper reference distribution for (5.6) is  $\chi^2_{n-p}$ .

Box [1980] found that many of the predictive checks for standard models were identical to hypothesis tests derived from corresponding sampling theory

models. This is also true of the current diagnostic statistic. An appropriate sampling theory model for (2.1) can be derived by viewing the distribution of  $Y$  given a fixed value for the regression vector  $\theta_2$ , rather than averaged over a prior distribution for  $\theta_2$ . The resulting model is a standard linear model, but with correlated errors:

$$(5.8) \quad Y \sim N(X\theta_2, \sigma^2 I_n + \lambda^{-1} \sigma^2 R).$$

Theorem 5.2: Given the sampling model (5.8) and assuming that  $\sigma^2$  is known, the appropriate lack of fit statistic for the model  $E\{Y\} = X\theta_2$  is (5.6). The sampling theory distribution of (5.6) is  $\chi^2_{n-p}$ . Thus Theorem 5.2 provides still further support that  $\chi^2_{n-p}$  is an appropriate reference distribution for (5.6).

The theory presented thus far might be used to test whether a particular value of  $\lambda$  is consistent with the observed data. It is common practice to apply reverse logic to such procedures and to consider the collection of  $\lambda$ 's which are consistent with the data, within given quantile bounds of the reference distribution. Traditionally, this is advocated to construct confidence sets from hypothesis tests. Whether or not we choose to regard the resulting collection of  $\lambda$ 's as a confidence set, the exercise is instructive as to the nature of the proposed predictive check. Theorem 5.2 can be used to characterize such a set, via the following corollary.

Corollary 5.2.1:  $h^*(\lambda)$  is a monotone increasing function of  $\lambda$ .

The corollary makes it clear that for any given quantiles from  $\chi^2_{n-p}$ , the set of  $\lambda$ 's for which  $h^*(\lambda)$  lies within the quantiles will be an interval (possibly infinite). This is intuitively reasonable. Moreover, we see that some choices of  $\lambda$  may be rejected because the diagnostic statistic is too large, while other choices may be deemed inconsistent with the data because the statistic is too small to reasonably have a  $\chi^2_{n-p}$  distribution. This

differs from the traditional lack of fit test, which rejects only in those instances where the statistic is too large. This is a consequence of the explicit representation of bias in the Bayesian models: the traditional test can tell us only that the approximating model is inadequate, while the predictive test tells us whether our representation of the inadequacy is insufficient, excessive or reasonable.

The predictive check (5.6) can also be extended to problems for which  $\sigma^2$  is unknown but can be independently estimated. Typically this occurs when replicate observations have been made at some of the design points -- since replicate observations have the same expected value, any differences among them can be attributed solely to experimental error. This leads to the decomposition of the residual sum of squares into "pure error" and "lack of fit" components given in (5.2). If  $m(i)$  observations,  $Y_{i,1}, \dots, Y_{i,m(i)}$ , have been made at  $x_i$ ,  $i=1,2,\dots,k$ , then the pure error sum of squares is given by:

$$(5.9) \quad SS(\text{Pure Error}) = \sum_{i=1}^k \sum_{j=1}^{m(i)} (Y_{i,j} - \bar{Y}_i)^2,$$

where  $\bar{Y}_i$  is the average of the  $m(i)$  observations at  $x_i$ . The lack of fit sum of squares can now be obtained by subtraction. Alternatively, we can calculate the lack of fit sum of squares by replacing each component of the response vector  $\mathbf{Y}$  by the average of all the replicate observations at that point and then comparing this vector to the predicted value vector. Specifically, let  $\bar{\mathbf{Y}}$  be the vector generated from  $\mathbf{Y}$  by replacing  $Y_{i,j}$  by  $\bar{Y}_i$ . Then the lack of fit sum of squares can be written:

$$(5.10) \quad SS(\text{Lack of Fit}) = (\hat{\mathbf{Y}} - \bar{\mathbf{Y}})'(\hat{\mathbf{Y}} - \bar{\mathbf{Y}}).$$

It is easy to show that (5.9) and (5.10) satisfy (5.2).

The diagnostic function  $h^*(\lambda)$  can also be decomposed into "pure error"

and "lack of fit" components when replicate observations are available. We define  $SS(\text{Pure Error})$  exactly as in (5.9), while  $SS(\text{Lack of Fit})$  is found by an approach analogous to that which led to (5.10). As above, let  $\bar{\mathbf{Y}}$  be a vector generated from  $\mathbf{Y}$  by replacing each observed response by the average of all the responses observed at the same design point. The lack of fit sum of squares is then calculated from (5.6), but with  $\bar{\mathbf{Y}}$  replacing  $\mathbf{Y}$ :

$$(5.11) \quad SS(\text{Lack of Fit}) = [\mathbf{B}(\lambda)\bar{\mathbf{Y}}]'(\mathbf{I}_n + \lambda^{-1}\mathbf{R})[\mathbf{B}(\lambda)\bar{\mathbf{Y}}].$$

It is easy to show that the sum of (5.9) and (5.11) is the weighted residual sum of squares in the numerator of (5.6).

From a sampling theory point of view, we can now test the adequacy of the model by comparing the ratio:

$$[SS(\text{Lack of Fit})/(k-p)]/[SS(\text{Pure Error})/(n-k)]$$

to quantiles of an F distribution with  $k-p$  and  $n-k$  degrees of freedom. We can also justify this procedure from a Bayesian standpoint by considering the distribution of this ratio conditional on  $\sigma^2$  and then assigning a uniform improper prior distribution to  $\log(\sigma)$ . If we assume that the distribution of  $h^*(\lambda)$ , conditional on  $\sigma^2$ , is actually  $\chi^2_{n-p}$ , then integrating out  $\sigma^2$  results in the F distribution above.

Although we have emphasized the use of  $h^*(\lambda)$  to check the consistency of  $\lambda$  with the data, it is important to keep in mind that it is the entire model specification that is being checked, not just the particular value of  $\lambda$ . Thus, the previous discussion really must be read as conditional on the assumption that we are quite certain of the best choice of an approximating polynomial and of the form of  $\mathbf{R}$ ; only the relative importance of bias to experimental error is really at question. If this is the case, then an abnormally large or small value of the diagnostic statistic can legitimately be interpreted as reflecting an inappropriate choice of  $\lambda$ , and we may seek to

improve the model by changing  $\lambda$ .

Often, there will be several aspects of the model about which we are unsure. Since the diagnostic statistic (5.6) is an omnibus check of the model, it will not point to any one of these features as inadequate. Rather, we must use our judgment in determining which aspects of the model should be changed in order to make it consistent with the data. Sometimes, a more flexible approximating polynomial may suffice to render the bias correction insignificant; this, of course, is the corrective action almost invariably used when the traditional lack of fit test is applied. Graphical techniques will often suggest when this is desirable. In other cases, revising our assumptions about the smoothness of the response function as reflected in  $R$  might be of great help. However, the predictive check can be most easily expressed in terms of its relation to  $\lambda$  and, in general, we feel it should be interpreted as an indication of the adequacy of  $\lambda$ .

## 6. A Kalman Filter Version

The Kalman filter is a system of stochastic equations used primarily by control engineers to describe processes that evolve over time in a structured, but non-deterministic fashion. The primary objective of the engineers is to study the properties of different strategies which might be used to keep the process under control. However, the Kalman filter itself is quite general and may be used to model a wide variety of statistical problems. A recent paper by Harrison and Stevens [1976] has suggested how the Kalman filter can be applied to the problem posed here of studying how a response variable depends on several inputs when an exact form for the response function is unknown. This section will show how the models of section 2 can be written as a Kalman filter; however, we will argue that the chief advantage of this model, Kalman's recursive estimation procedure, cannot be applied.

We will follow the general set-up of Harrison and Stevens. They proposed a "dynamic linear model":

$$(6.1a) \quad \mathbf{y}_t = \mathbf{F}_t \boldsymbol{\theta}_t + \mathbf{u}_t, \quad \mathbf{u}_t \sim N(0, \mathbf{U}_t)$$

$$(6.1b) \quad \boldsymbol{\theta}_t = \mathbf{G} \boldsymbol{\theta}_{t-1} + \mathbf{w}_t, \quad \mathbf{w}_t \sim N(0, \mathbf{W}_t).$$

The first equation, known to control engineers as the observation equation, relates a vector of observed variables,  $\mathbf{y}_t$ , to a known matrix  $\mathbf{F}_t$  of regressor variables via a linear model plus random noise. The second, or system, equation relates how the parameters themselves may change from one observation to the next (herein lies the "dynamic" element of the model). It is assumed that the current parameter values are a known linear transformation  $\mathbf{G}$  of the previous values perturbed by the random error  $\mathbf{w}_t$ . Initial parameter estimates,  $\boldsymbol{\theta}_0$ , are needed to start the process. Typically, rather than specify exact values, a prior distribution is provided for  $\boldsymbol{\theta}_0$ , so that the model has a distinctly Bayesian character.

Harrison and Stevens proposed this model with time series data in mind, so they naturally interpreted the index  $t$  as denoting the sequential order of the observations. However, the model is not restricted to time dependent data. In general, one can think of  $t$  as simply an arbitrary parameter which labels the different observations (perhaps the order in which experimental runs were made, or the "standard" ordering for the experimental design used). The matrix  $F_t$  then corresponds only to the values of the regressor variables for the  $t$ 'th observation, not the entire  $X$  matrix of the previous sections.

The rationale for ordering the observations is so that we may use Kalman's recursive estimation algorithm. The algorithm provides a simple but powerful way to calculate the posterior distribution of  $\theta_t$  given  $y_1, \dots, y_t$  and  $F_1, \dots, F_t$ . Specifically, if the prior distribution of  $\theta_0$  is  $N(m_0, C_0)$ , then:

$$(6.2) \quad \theta_t / y_1, \dots, y_t, F_1, \dots, F_t \sim N(m_t, C_t),$$

where  $m_t$  and  $C_t$  are obtained recursively from the following set of equations. Let:

$$(6.3a) \quad y = F_t G_{t-1},$$

$$(6.3b) \quad e = y_t - y,$$

$$(6.3c) \quad R = G C_{t-1} G' + W_t,$$

$$(6.3d) \quad S = F_t R F_t' + V_t,$$

$$(6.3e) \quad A = R F_t' S^{-1}.$$

Then:

$$(6.4a) \quad m_t = G_{t-1} + A e,$$

$$(6.4b) \quad C_t = R - A S A'.$$

Note that if the  $t$ 'th observation,  $y_t$ , is  $p$ -dimensional, then equations (6.3) and (6.4) provide a technique for calculating the posterior distribution of

the parameter vector  $\theta_t$  which never requires inversion of matrices larger than  $p \times p$ , regardless of the sample size. In particular, if  $y_t$  is a scalar, then no matrix inversion is required at all. This special case was remarked by Plackett [1950] as a simple device for updating regression estimates when a new data point is obtained.

Many common statistical models can be written in the form of (6.1). For example, standard linear regression can be described by taking  $G$  to be the identity matrix and requiring that  $w_t$  be identically 0 for each  $t$ ; that is, each successive observation depends in exactly the same way on the regressors. Harrison and Stevens present numerous additional examples in their paper. In particular, they remark that "hierarchical models, such as those of Lindley and Smith [1972], can also be formulated as dynamic linear models". The Kalman filter approach is clearly evident in O'Hagan's [1978] localized regression model, which postulates a regression model in which the parameters change regularly from one point in the design space to another (see in particular the comments by Priestley and Titterton).

We will now show how Smith's model (2.1), itself a special case of the models proposed by Lindley and Smith, can be written as a dynamic linear model. (However, we will use the equivalent formulation of Blight and Ott ((2.3) and (2.4)) rather than (2.1) in order to avoid confusion with respect to the indices on  $\theta$  -- in (2.1) the indices 1 and 2 are used to denote parameters at the first and second tiers of the model, respectively, while in (6.1) the same indices are used to denote the regression parameters for the first and second observations, respectively.)

Recall from (2.3) that the vector of observed responses follows the model:  $Y = X\beta + \eta + \epsilon$ . Let  $t$  be an arbitrary label, as suggested above. Then the model equation for the  $t$ 'th observation is:

$$(6.5a) \quad Y_t = z_t' \beta + \eta_t + \epsilon_t = z_t' \beta + N_t,$$

where  $N_t = \eta_t + \epsilon_t$ . Since we wish to describe (2.3) on an observation by observation basis,  $Y_t$  is a scalar, rather than the more general vector observation allowed in (6.1). From the distributional assumptions for the Blight and Ott model (2.4), we conclude that  $N_t$  has a normal distribution with mean 0 and variance  $\sigma^2 + l_t$ , where  $l_t$  is the appropriate diagonal entry from  $L^{-1}$ . Thus, (6.5a) gives the observation equation for the model.

The system equation which corresponds to (2.3) is derived in the same manner mentioned above for a conventional linear regression model. It is assumed in (2.3) that the regression parameters remain the same for each observation. This implies the system equation:

$$(6.5b) \quad \beta_t = \beta_{t-1},$$

where  $G_t = I$  and  $w_t = 0$  for each  $t$ . The prior distribution for  $\beta$  given in (2.4) can then be introduced by assuming that it is the prior distribution for  $\beta_0$ . Since the system equation implies that  $\beta_0 = \beta_1 = \dots = \beta_n$ , the same prior distribution will be in effect for each observation. Thus for each observation, the Blight-Ott model, and hence the Smith model, can be written as a dynamic linear model.

The great advantage of the Kalman filter model is the existence of the recursive estimation algorithm (6.2)-(6.4) which provides a simple method to calculate all the relevant posterior distributions without having to invert large matrices. Unfortunately, the Kalman recursion relations are not valid for (6.5). The reason is an important assumption, unstated by Harrison and Stevens in either (6.1) or (6.2), that the random shocks occurring at different time points must be uncorrelated; i.e., that if  $t \neq t'$ , then  $(u_t, w_t)$  must be uncorrelated with  $(u_{t'}, w_{t'})$ . However, we found in section 3 that the assumption that  $u_t$  and  $u_{t'}$  are uncorrelated for distinct  $t, t'$  leads to a

discontinuous estimator. Thus the assumption that they are in fact correlated is a central feature of the model. The algorithm can be extended to cover this case (see Harrison and Stevens' reply to Godolphin's comment), but the extension requires the inversion of a  $j \times j$  matrix at the  $j$ 'th step in order to obtain the conditional distribution of  $v_j$  given  $v_1, \dots, v_{j-1}$ . Moreover, this provides only the relevant posterior distributions for the observed sequence of data points; in order to draw predictive inferences about possible future observations, it is still necessary to invert an  $n \times n$  matrix.

The Kalman filter representation provides an interesting alternative statement of the models, emphasizing the similarity of the Bayesian linear model and control theory models. However, unless the response function is naturally observed as a time series, it is questionable whether the Kalman filter representation provides much additional insight; it seems, in fact, to be a rather awkward way to think of general response surface models. Nonetheless, the representation would certainly be very useful if the computational efficiency of Kalman's recursive algorithm were applicable; it is unfortunate that this is not so.

## 7. Examples

In this section we will illustrate the results of the previous sections by using the Bayesian models to analyze two numerical examples. Both examples involve the response to a single input, and in both cases the response is essentially a linear function of the input over the range the observations. Thus, the straight line approximating model (1.1a) will be used, with the input as the only explanatory variable. We will assign a diffuse prior distribution to the regression parameters, so that the "standardized" notation of Corollary 3.2.2 is appropriate. All of the calculations described in this section were performed on the WIRCS/VAX computer. The Bayesian estimates were calculated using the MATLAB matrix laboratory package (Moler, [1981]) while the standard regression analyses were carried out using the MINITAB statistical package (Ryan, et. al. [1981]).

The first example uses data reported by Draper and Smith [1981] from an experiment to investigate the relationship between the yield of a chemical process and the reaction temperature (in coded units):

Temp	-5	-4	-3	-2	-1	0	1	2	3	4	5
Yield	1	5	4	7	10	8	9	13	14	13	18

A casual inspection of the data is sufficient to conclude that yield generally increases as a function of temperature. The increase across the range of temperatures in the experiment seems to be more or less linear, although there is a fair amount of variation from one observation to the next, suggesting either that experimental error may be fairly substantial, or that some additional explanatory variable(s) which was not included in the experiment is important in determining yield.

We propose to model these data by (2.3):

$$Y = X\beta + \eta + \epsilon,$$

where  $Y$  is the vector of observed yields,  $X$  is the design matrix which corresponds to a linear approximating model in temperature,  $\eta$  is the vector of bias terms and  $\epsilon$  is the vector of experimental errors. We will assume that the autocorrelation function of the bias process is:

$$(7.1) \quad R(x_1, x_2) = \exp[-(x_1 - x_2)^2/s],$$

where  $s$  is a parameter which we will specify. This is a legitimate autocorrelation function (by Bochner's Theorem -- see Chung [1974], Chapter 6) and corresponds to a continuous, second order, stationary Gaussian stochastic process. Note that  $R(x, x) = 1$  for all real  $x$ ; thus, we are assuming in (7.1) that the probable extent of the bias is the same at all temperatures. The choice of  $s$  controls how rapidly the correlation drops off between different temperatures. If  $s$  is quite small, the decay is very rapid; this corresponds to a prior belief that the relationship of yield to temperature may have sharp local deviations from the overall linear approximation. On the other hand, if  $s$  is large, so that there is a high correlation between the bias terms for rather different temperatures, then (7.1) will reflect prior belief that the response does not have any wild local oscillations. Thus the choice of  $s$  is closely linked to the experimenter's prior beliefs about how smoothly yield responds to temperature.

The autocorrelation function in (7.1) is very similar to that proposed by Blight and Ott:  $R(x, z) = \lambda^{|x-z|}$ ,  $0 < \lambda < 1$ . The only difference is the use of the squared difference rather than the absolute difference between the respective points. However, a possibly important consequence of this change is that it results in a prediction equation which is an analytic function of temperature rather than a function with a discontinuous first derivative, as results from the Blight-Ott suggestion.

Two different values were chosen for each of the parameters associated with the bias covariance function. The parameter  $s$  in the correlation function was set at 1 and 16, so that the correlation decays to  $1/e$  for temperatures which are 1 or 4 coded units apart, respectively. The experimental error to bias parameter  $\lambda$  was set at 1 and 0.1, reflecting situations in which bias is of the same magnitude or ten times as severe as experimental error, respectively. Using each combination of these parameter settings gave four different prediction equations. A fifth prediction equation was derived by using ordinary least squares. The parameter settings were chosen merely to illustrate the flexibility of the Bayesian models and to illustrate their application; they do not reflect any experimenter's prior assumptions about the relationship between temperature and yield.

The four prediction functions described above are graphed, along with the OLS prediction function, in the four panels of Figure 1; each graph also displays the observed data points. Overall, the five functions are very similar and are clearly dominated by a positive linear trend. Over the range of values graphed, these functions never differ by more than a unit of yield. In particular, an analyst using OLS regression would reach roughly the same conclusions about the relationship between yield and temperature as a colleague using any of the suggested Bayesian models.

The differences among the prediction functions concern primarily the manner in which they vary about the linear trend. The most noticeable differences occur when  $s$  is altered from 1 to 16, reflecting prior belief in a smoother relationship. The prediction functions bear convincing witness to the effect of such a prior belief: the two graphs with  $s=1$  display considerable local variation, with fairly sharp increases in yield followed by plateaus and even by decreases. Although these prediction functions follow an

$S=1 \quad \lambda=1$

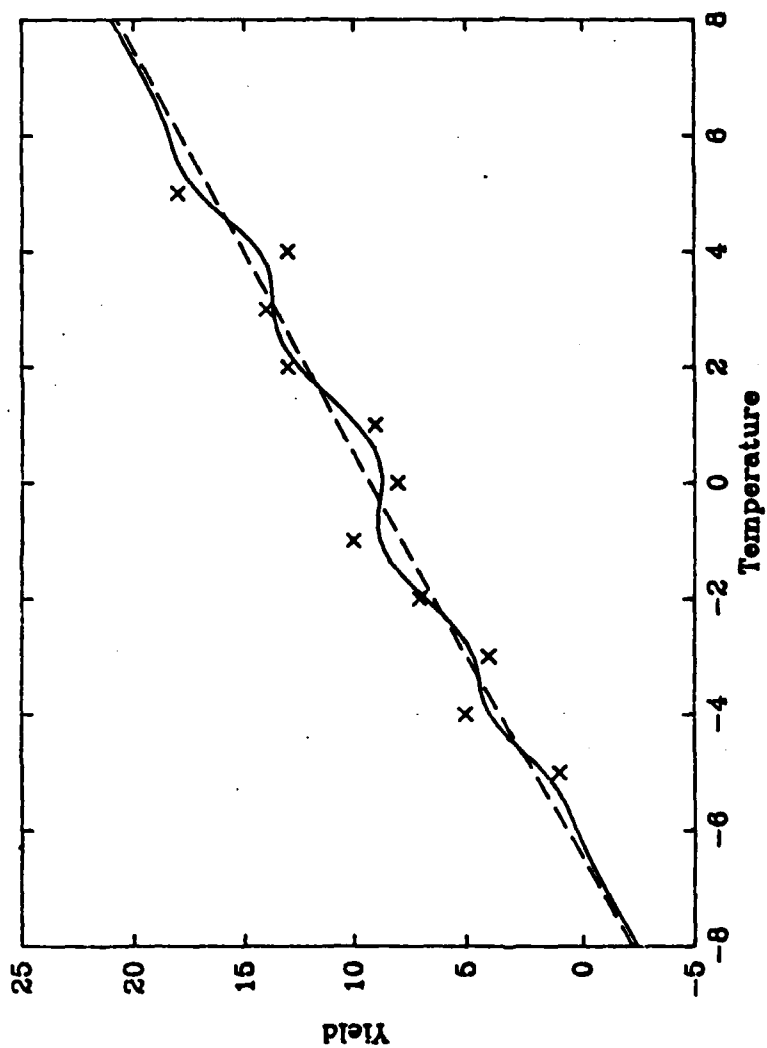


Figure 1(a): The Bayesian prediction function for  $s=1, \lambda=1$  (solid line) and for OLS regression (dashed line) for the Draper and Smith data. The data points are marked by x's.

$S=1 \quad \lambda=0.1$

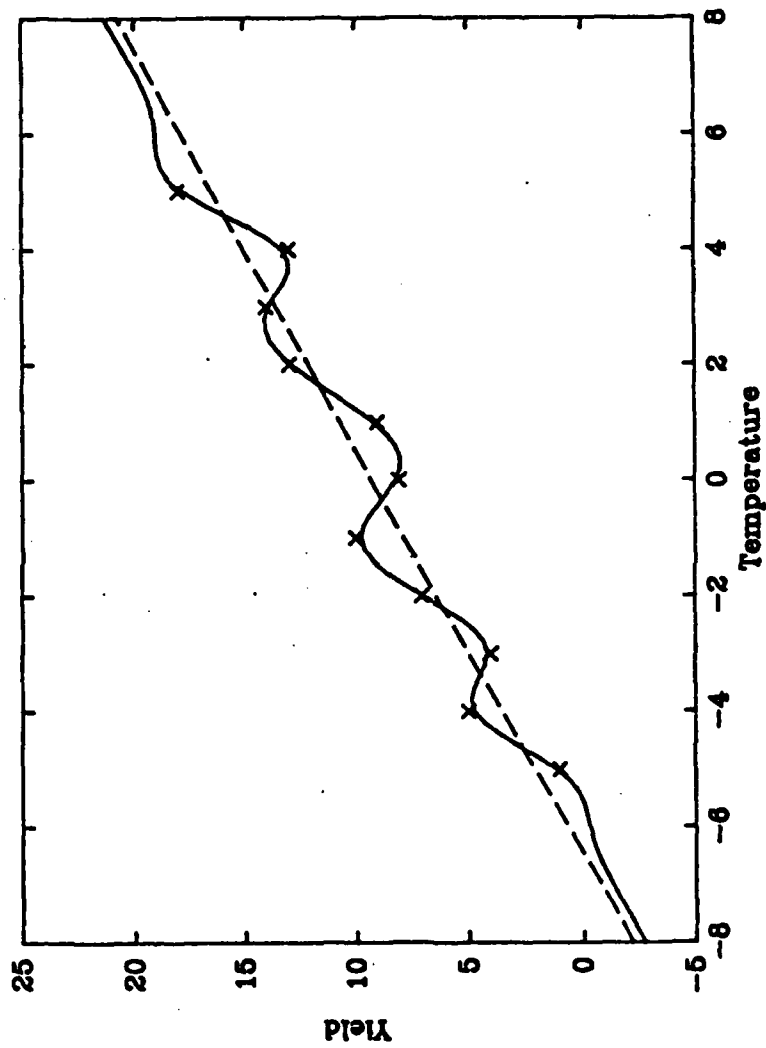


Figure 1(b): The Bayesian prediction function for  $s=1$ ,  $\lambda=0.1$  (solid line) and for OLS regression (dashed line) for the Draper and Smith data. The data points are marked by x's.

$S=16 \quad \lambda=1$

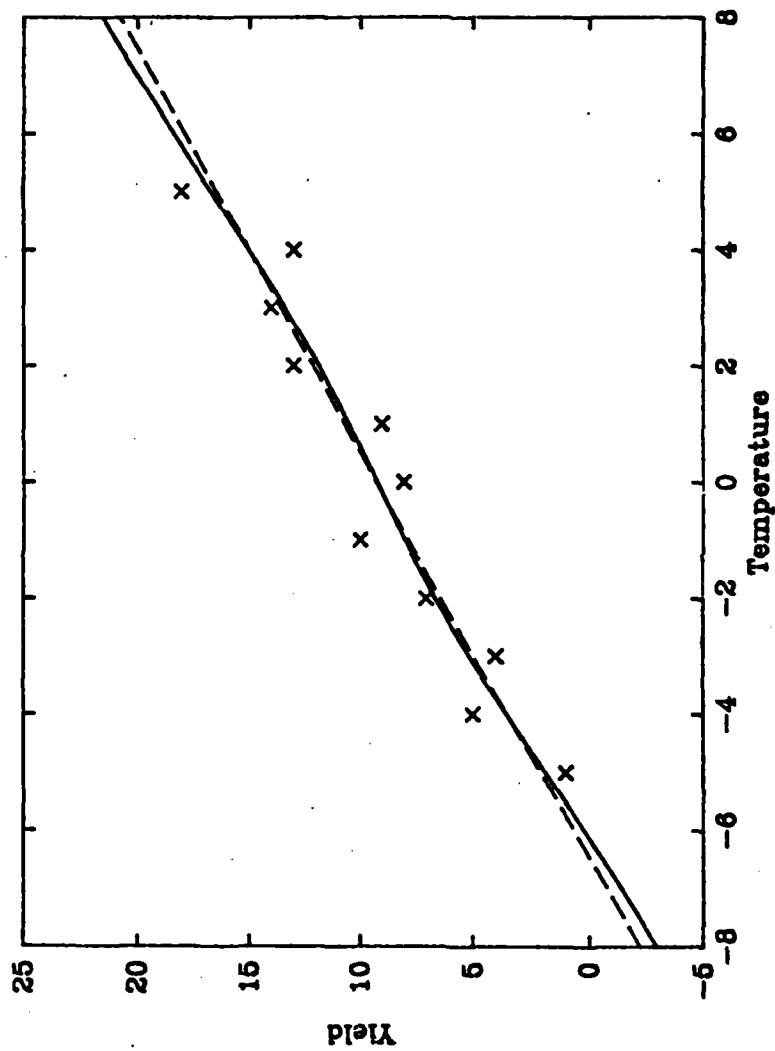


Figure 1(c): The Bayesian prediction function for  $s=16$ ,  $\lambda=1$  (solid line) and for OLS regression (dashed line) for the Draper and Smith data. The data points are marked by x's.

$S=16 \quad \lambda=0.1$

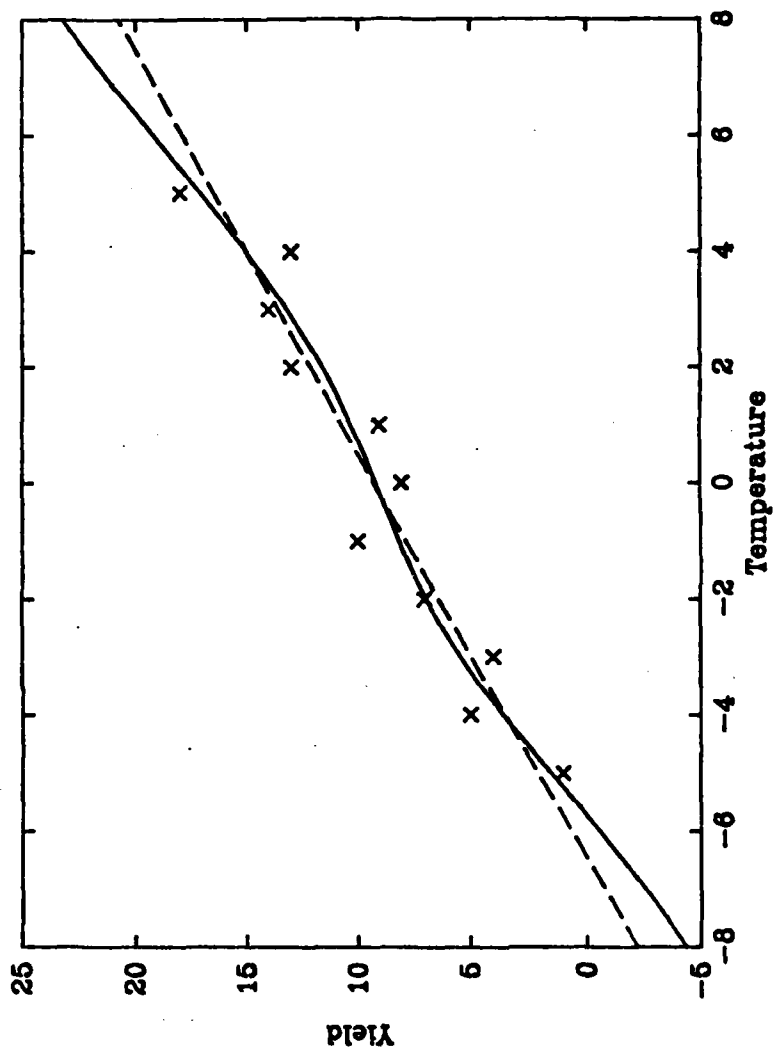


Figure 1(d): The Bayesian prediction function for  $s=16$ ,  $\lambda=0.1$  (solid line) and for OLS regression (dashed line) for the Draper and Smith data. The data points are marked by X's.

overall linear increase, their slopes vary considerably with temperature. By contrast, the two graphs with  $s=16$  are strictly monotone increasing and their slopes remain close to the overall trend at all temperatures.

The experimental error to bias tradeoff parameter  $\lambda$  is seen to have less effect than the correlation parameter  $s$  in this example. The two curves for  $s=16$ , in particular, are almost identical throughout the region of the observations. Only for those temperatures where our predictions are extrapolations from the experimental region are there any noticeable differences. When  $s=1$ , changing  $\lambda$  has a more dramatic effect. The prediction equation for  $s=1$ ,  $\lambda=0.1$  follows the observed data very closely; at each of the experimental temperatures its predictions are closer to the observed yields than any of the other equations. In fact, this equation seems to follow the data "too well"; it would probably be very difficult to convince the experimenter that the effect of increasing temperature on yield is so irregular. It is more likely that we will be convinced that our parameter settings are unrealistic, exaggerating the effect of bias in shifting the observed yields off a straight line when experimental error may have been much more influential. We know from Theorem 3.3 that in the limit, as  $\lambda \rightarrow 0$ , the prediction equation will exactly interpolate the data (since we are using a non-singular  $R$  here), and it is clear that for  $s=1$ , decreasing  $\lambda$  to 0.1 is sufficient to come very close to an interpolant. With  $s=16$ , even this ten to one bias to error ratio is still a long way from interpolating the data. Thus we also see, at least in this example, an "interaction" between the two parameters.

The estimated regression parameters and their standard deviations (divided by  $\sigma$ ) are given in Table 1. The parameter estimates take on an added importance in light of the particular bias covariance function we are

using. Note that  $R(x_1, x_2) \rightarrow 0$  as  $|x_1 - x_2| \rightarrow \infty$ . It follows from (3.5) that for this example the prediction equations will converge asymptotically to the straight lines determined by the parameter estimates. Thus, even if the data had given rise to prediction equations that differ considerably from an overall linear trend in the experimental region, the form of the bias covariance function guarantees that extrapolation to points far enough outside that region will be linear in temperature.

The estimated regression parameters are similar for all the models with the exception of  $s=16, \lambda=0.1$ . This bias covariance specification estimates a slope which is more than 10% larger than that for any of the other models and more than 17% larger than the OLS estimate. One possible explanation for this involves the effect of the most distant observations on extrapolation when the bias covariance is assumed to be rather stable. Intuitively, if the yield corresponding to the highest temperature is "unusually" high, and we think that the bias at proximate temperatures is quite similar, then this highest observation may play a disproportionate role in determining our extrapolations; mathematically, this is evident from examining the terms in (3.5). Such an effect will be strengthened the more we decrease  $\lambda$ ,

Table 1: The estimated regression parameters and their standard deviations (divided by  $\sigma$ ) for the four settings of  $\lambda$  and  $s$  and for OLS regression.

	Constant	(S.D./ $\sigma$ )	Slope	(S.D./ $\sigma$ )
$\lambda=1, s=1$	9.282	(0.49)	1.468	(0.15)
$\lambda=0.1, s=1$	9.277	(1.27)	1.508	(0.38)
$\lambda=1, s=16$	9.288	(0.75)	1.516	(0.19)
$\lambda=0.1, s=16$	9.341	(2.12)	1.682	(0.46)
OLS	9.273	(0.30)	1.436	(0.10)

emphasizing bias to a greater degree. But we noted above that extrapolations are also related to the estimated regression coefficients. In this example, the two most extreme temperatures have observed yields which suggest a steeper slope than do the remainder of the observations. This may account for the fact that all the Bayesian estimates of the slope are greater than the OLS estimate. This effect becomes dramatic when  $s=16$ , reflecting prior belief in a "stable" bias contribution, and  $\lambda=0.1$ , indicating a belief that bias greatly dominates experimental error in determining observed deviations from a straight line graduating function.

The estimated standard deviations of the regression coefficients, unlike the coefficients themselves, differ considerably from one model to the next. Here it is  $\lambda$  which has the principal effect. Decreasing  $\lambda$  from 1 to 0.1 increases the standard deviation of both coefficients (modulo  $\sigma$ ) by over 2.5 times. Increasing  $s$  also increases the standard deviations, but to a more modest extent. The effect of  $\lambda$  is really a logical consequence of the prior assumptions of the model: if a straight line is our best guess as to the overall dependence of yield on temperature, but we believe there may be rather severe bias, then it only seems reasonable that the experiment will be unable to provide us with precise estimates of the coefficients of the straight line. However, an additional point should also be kept in mind: if we believe that bias dominates experimental error, then for any given data set we will presumably obtain a lower estimate of  $\sigma$  which will offset some of the differences evident in the table.

The Bayesian models lead to quite different conclusions than OLS about our ability to predict yield as a function of temperature. Figure 2 displays graphs of  $\text{Var}\{Y_x/Y-\bar{Y}\}/\sigma^2$  for (7.1) with  $s=1$  and 16 and  $\lambda=1$ , and for OLS. These functions are clearly symmetric about 0, so the graphs extend only over

non-negative temperatures. It is clear from the graphs that OLS is more optimistic about our ability to predict yield than are the Bayesian models. This is perfectly reasonable, since the Bayesian models provide for a greater degree of prior uncertainty about the experiment. The two Bayesian models shown in Figure 2 also differ considerably. The prediction variances with  $s=16$  are much lower than with  $s=1$ . Again this is a logical consequence of the differing prior beliefs being modeled. If  $s=16$ , the bias is relatively stable, and the experiment permits us to effectively ascertain its effect at least throughout the experimental region; however, if  $s=1$ , we are asserting a prior belief that the bias may exhibit considerable local fluctuation so that only observations in the immediate vicinity of a given temperature will really be of much use in predicting yield there. This is the exact antithesis of OLS, where certainty about the linear form of the response function implies that experimenting at the most extreme temperatures possible will lead to the most precise predictions at all temperatures. It also explains the local minima that we observe in the graph for  $s=1$  near most of the actual experimental temperatures, while the other graphs are strictly monotone increasing.

Within the range of the experimental data, the Bayesian model with  $s=16$  produces prediction variances about 7% higher than OLS while the model with  $s=1$  produces variances about 40% higher. However, recall that these are the variances of predicting a new observation and thus they all include a contribution of 1 for the experimental error associated with that observation. If we subtract that term to obtain variances for the associated expected value, the increases in the variances over OLS are about 40% and 250%, respectively. Outside the range of the data the differences are still greater. The prediction variances for the Bayesian models increase rapidly

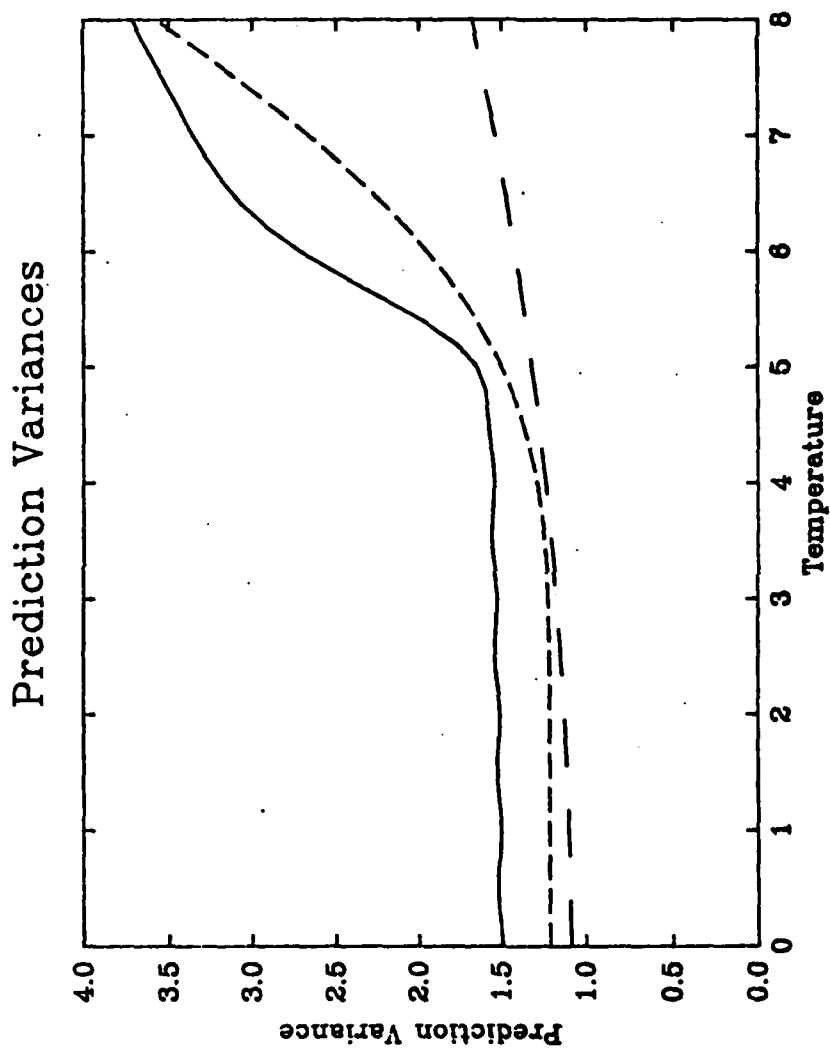


Figure 2: Prediction variance,  $\text{Var}\{Y_x/Y=y\}/\sigma^2$ , as a function of temperature for the Draper and Smith data, for  $s=1$ ,  $\lambda=1$  (solid line), for  $s=16$ ,  $\lambda=1$  (short dashes), and for OLS regression (long dashes).

for extrapolating while the OLS variances continue to increase quadratically.

The second example involves an analysis of simulated data. Suppose the true, but unknown response function is:

$$(7.2) \quad g(x) = 20[\exp(-0.3x) - \exp(-0.4x)].$$

This response function is graphed in Figure 3. Response functions of this form often arise in chemical reactions when an initial product A decays to an intermediate product B which then decays to a final product C. Suppose an experiment is conducted to investigate the behaviour of the response in the neighborhood of  $x=5$ , where it is known that the response is approximately linear in  $x$ . The experiment calls for taking observations at 11 equally spaced sites from  $x=2.5$  to  $x=7.5$ . We will assume that  $\sigma^2$  is known in advance to be 0.01.

We propose to model the observed data by  $Y = X\beta + \eta + \epsilon$ , where  $X$  is the design matrix for straight line regression,  $\eta$  is the bias term and  $\epsilon$  the vector of experimental errors. We will model the bias covariance matrix by:

$$(7.3a) \quad R(x_1, x_2) = [(x_1 - 5)(x_2 - 5)/4] \exp(-(x_1 - x_2)^2/2) \\ \text{if } (x_1 - 5)(x_2 - 5) > 0$$

$$(7.3b) \quad R(x_1, x_2) = 0 \quad \text{otherwise.}$$

The idea behind this covariance function is to represent prior belief that a straight line approximation is appropriate in the vicinity of  $x=5$ , but may be increasingly suspect as  $x$  becomes more distant from 5. This sort of assumption is appropriate for many response surface experiments. Thus the bias variance at  $x$  is equal to  $(x-5)^2/4$ . The division by 4 is a standardization device, so that the choice of  $\lambda$  will reflect the experimental error to bias ratio at  $x=3$  and  $x=7$ , covering 80% of the experimental region. We will suppose that at these points bias is of about the same magnitude as experimental error so that  $\lambda=1$ . The second term in (7.3a) reflects a prior

# True Response Function

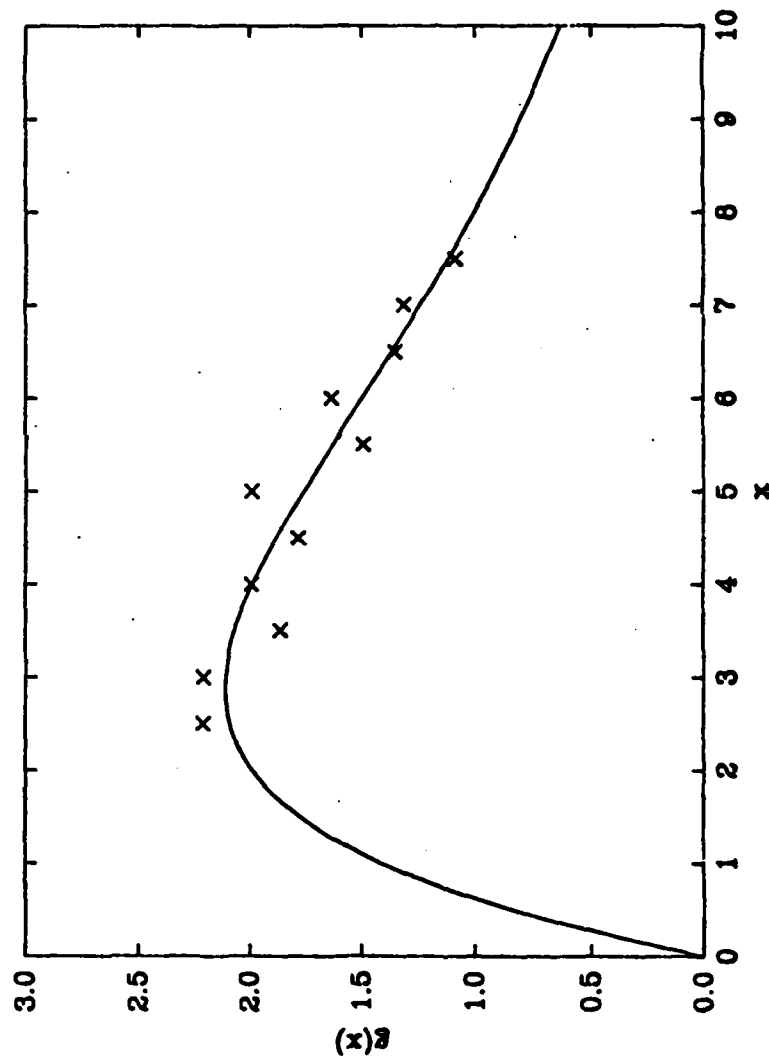


Figure 3: The true, but unknown, response function  $g(x)=20[\exp(-0.3x) - \exp(-0.4x)]$  (solid line) and the "experimental data" (x's) for the second example.

belief that the covariance between the bias contributions at various  $x$ 's tends to 0 as the  $x$ 's become increasingly distant from one another. The denominator under the square in the exponential term controls the rate at which this happens and has been chosen to reflect a reasonable belief about the smoothness of the response function. A possible drawback to (7.3) is that for any given  $x_1$ ,  $R(x, x_1)$  is a continuous function of  $x$ , but its derivatives are discontinuous at  $x=5$ . Thus the derivatives of the prediction equation will also be discontinuous at  $x=5$ . Also, note that  $R(5, x)=0$  for all  $x$ , and that  $x=5$  is one of the observation sites. Thus the  $R$  matrix for this model will be singular, so that some of the results of Theorems 3.3 and 3.4 will not apply to this example.

To justify that (7.3) is a legitimate covariance function, consider two independent stochastic processes, each beginning at  $x=5$  but moving from there in opposite directions on the  $x$  axis. Suppose that each process, on those  $x$ 's for which it is defined, has the covariance function given by (7.1) with  $s=2$ . Define two new stochastic processes by multiplying each of the original processes by  $(x-5)/2$ . Since both of the new processes are constrained to be deterministically 0 at  $x=5$ , we may tie them together into a single stochastic process defined on the entire real line. This new stochastic process is easily seen to have (7.3) as its covariance function.

The "experimental data" for this example were obtained by calculating the actual function values at the design points from (7.2) and then adding to them computer generated random errors. The random errors were generated using the normal distribution random number routine in MINITAB (Ryan, et. al. [1981]). The function values and the observations are listed in Table 2. The observations are also included on the graph of  $g(x)$  in Figure 3.

The Bayesian prediction equation and the OLS prediction line are both

graphed in Figure 4. They are very similar to one another, although between 7.5 and 8.5 there is a mild discrepancy. This evidently results from the effect of the low observation at 7.5 on the Bayesian line, in much the same manner as we discussed for the last example. The estimated regression coefficients and their standard errors are listed in Table 3. The standard errors were calculated using the assumed known value  $\sigma^2 = 0.01$ . The estimates are almost identical, but as in the previous example, the standard errors are markedly greater with the Bayesian model.

Since  $\sigma^2$  is assumed to be known here, we may use the predictive model check described in section 5. The weighted residual sum of squares is 0.1482, and dividing by  $\sigma^2$  we obtain the checking statistic 14.82. The upper 10%

Table 2: The true response function values  $g(x)$  and the simulated data points  $Y$  for the second example.

<u>x</u>	<u>g(x)</u>	<u>Y</u>
2.5	2.0897	2.2079
3.0	2.1075	2.2087
3.5	2.0668	1.8612
4.0	1.9860	1.9903
4.5	1.8788	1.7798
5.0	1.7559	1.9895
5.5	1.6249	1.4901
6.0	1.4916	1.6316
6.5	1.3600	1.3449
7.0	1.2329	1.3037
7.5	1.1122	1.0837

## Predicted Response Curves

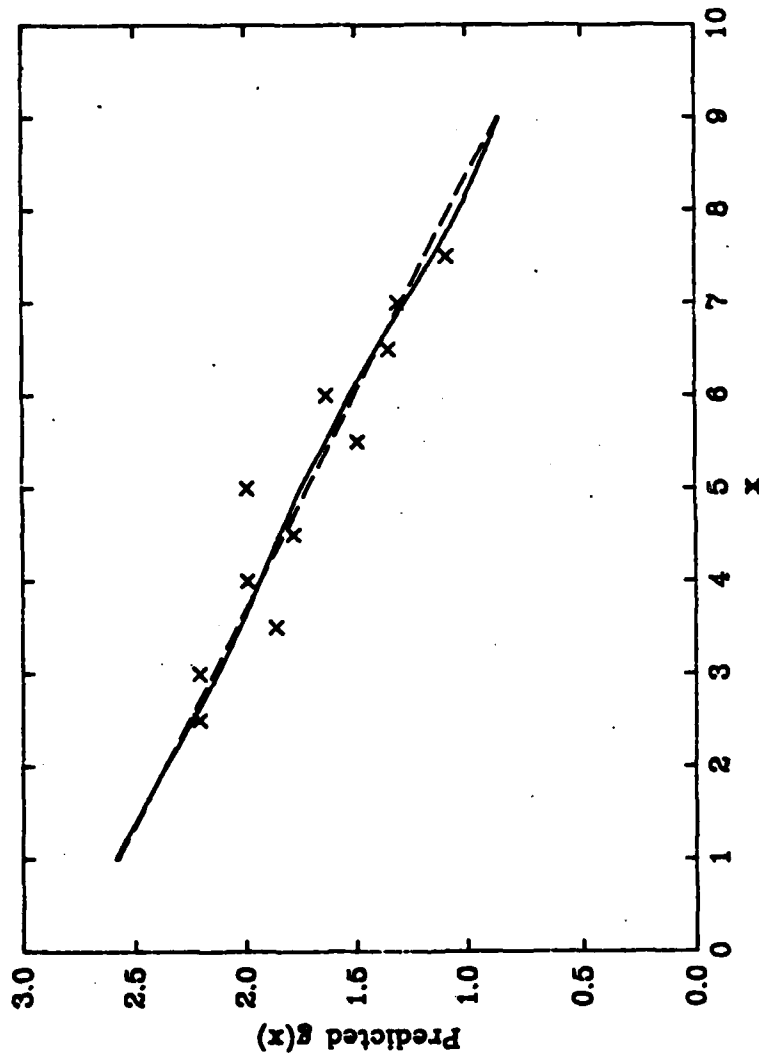


Figure 4: The Bayesian prediction function (solid line) and the OLS regression prediction function (dashed line) for the second example. The data points are marked by x's.

point of the  $\chi^2_9$  distribution is 14.68, so that there is about a 20% chance, if the model is valid, of obtaining so extreme a value for the checking statistic. This may raise some doubts about the proposed model, but is certainly not convincing proof of model inadequacy. Moreover, closer inspection reveals that the simulated random errors are "a bit large" for the  $N(0,0.01)$  distribution that was used -- the sample standard deviation of the errors is 0.13 rather than 0.10 -- and this has helped to inflate the checking statistic. Since other residual checks do not point to any gross model inadequacies, it seems reasonable to accept the validity of the Bayesian analysis.

Table 3: The estimated regression parameters and their standard deviations for the Bayesian model and for OLS regression.

	Constant	(S.D.)	Slope	(S.D.)
Bayesian model	2.815	(0.188)	-0.213	(0.037)
OLS regression	2.790	(0.100)	-0.214	(0.019)

## 8. Conclusions

Box [1982] observes that "all models are wrong; some models are useful. This aphorism must be particularly true of empirical models such as polynomials that make no claim to do more than locally approximate the true function." This paper has presented and analyzed several suggestions to modify standard response surface models to take into account the inherently approximate nature of commonly used graduating functions: the hierarchical model of Smith [1973], the "polynomial approximation + bias" model of Blight and Ott [1975], the "localized regression model" of O'Hagan [1978] and the smoothing spline approach of Wahba [1978].

The major result of the paper has been to demonstrate that these four models are essentially equivalent. This has helped make possible a complete analysis of the consequences of these models for estimating response surfaces, synthesizing and expanding upon the results which had been proved for each individual model. The estimates depend on the magnitude of bias relative to experimental error, with ordinary least squares regression obtaining in the "special case" when it is assumed that there is no bias at all. A predictive check of the model has been developed and is especially useful for criticizing assumptions about the ratio of bias to experimental error. Additionally, the relationship between these models and the Kalman filter has been explored. Although the models can be written as a Kalman filter, this does not seem to be a useful approach for most response surface models.

Much effort has been devoted in recent years to developing robust statistical procedures, that is, statistical procedures which will give reliable answers even when assumptions about the experimental mechanism (i.e., the proposed statistical model) are inexact. It is useful to view the models discussed in this paper in the context of the robustness literature. All four

models attempt to provide more reliable answers, but not by altering our methods to make them less sensitive to faulty models; rather, by recognizing the approximate nature of empirical graduating functions, they strive to provide a more realistic model whose assumptions more accurately reflect what we actually believe about the experimental data observed. This is consistent with the viewpoint advocated by Box [1980], who characterizes robustification as the "judicious and grudging elaboration of the model to ensure against particular hazards". The models provide a useful and flexible method for studying the adequacy of empirical response functions.

#### Acknowledgements

I am indebted to Professor George Box for suggesting this area of research to me and for his valuable insights and encouragement throughout.

## Appendix

We now prove the various distributional results stated in sections 3, 4 and 5. Most of the proofs rely on the following lemma, which gives the joint distribution of the observed responses, the regression parameters and a future (as yet unobserved) response at  $\mathbf{x}$ .

**Lemma 3.1:** Let  $\mathbf{Y}$  denote the response vector,  $Y_{\mathbf{x}}$  a future response at  $\mathbf{x}$  and  $\theta_2$  the regression parameters. Then, under the model described by (3.1), the joint probability distribution of  $(\mathbf{Y}', Y_{\mathbf{x}}, \theta_2')$  is multivariate normal with mean vector:

$$(A.1) \quad E\{(\mathbf{Y}', Y_{\mathbf{x}}, \theta_2')\} = ((\mathbf{X}\beta_0)', (z'\beta_0)', \beta_0')$$

and variance-covariance matrix:

$$(A.2) \quad \text{cov}\{(\mathbf{Y}', Y_{\mathbf{x}}, \theta_2')\} = \begin{vmatrix} \sigma^2 \mathbf{I}_n + \mathbf{V} + \mathbf{XV}_1\mathbf{X}' & \mathbf{v} + \mathbf{XV}_1\mathbf{z} & \mathbf{XV}_1 \\ \mathbf{v}' + \mathbf{z}'\mathbf{V}_1\mathbf{X}' & \sigma^2 + \mathbf{v} + \mathbf{z}'\mathbf{V}_1\mathbf{z} & \mathbf{z}'\mathbf{V}_1 \\ \mathbf{V}_1\mathbf{X}' & \mathbf{V}_1\mathbf{z} & \mathbf{V}_1 \end{vmatrix}.$$

**Proof:** By Theorem 2.1, we may write the hierarchical model (3.1) in terms of the Blight-Ott formulation (2.3)-(2.4):

$$\mathbf{Y} = \mathbf{X}\theta_2 + \eta + \epsilon,$$

$$Y_{\mathbf{x}} = z'\theta_2 + \eta_{\mathbf{x}} + \epsilon_{\mathbf{x}},$$

where  $(\eta', \eta_{\mathbf{x}})' \sim N(\mathbf{0}, \mathbf{V}^*)$ ,  $(\epsilon', \epsilon_{\mathbf{x}})' \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{n+1})$ , and  $\theta_2 \sim N(\beta_0, \mathbf{V}_1)$ , and where these three random vectors are independent. Equations (A.1) and (A.2) now result from simple and straightforward computation.

**Theorem 3.1:**

$$(3.2) \quad E\{Y_{\mathbf{x}}/\mathbf{Y}=\mathbf{y}\} = z'\beta_0 + (\mathbf{v}' + \mathbf{z}'\mathbf{V}_1\mathbf{X}')(\sigma^2 \mathbf{I}_n + \mathbf{V} + \mathbf{XV}_1\mathbf{X}')^{-1}(\mathbf{y} - \mathbf{X}\beta_0).$$

$$(3.3) \quad E\{\theta_2/\mathbf{Y}=\mathbf{y}\} = \beta_0 + \mathbf{V}_1\mathbf{X}'(\sigma^2 \mathbf{I}_n + \mathbf{V} + \mathbf{XV}_1\mathbf{X}')^{-1}(\mathbf{y} - \mathbf{X}\beta_0).$$

**Proof:** These equations follow directly from Lemma 3.1 by applying standard formulas for conditional expectations of partitioned multivariate normal vectors (see, for example, Anderson [1958], p. 28). The only condition which

must be checked in order to apply these formulas is that the covariance matrix in (A.2) is non-singular. Equivalently, we can show that for any non-zero vector  $a' = (a'_1, a'_2, a'_3)$ , the random variable  $(Y', Y_x, \theta'_2)a$  does not have a deterministic distribution. But

$$\begin{aligned}(Y', Y_x, \theta'_2)a &= a'_1 Y + a'_2 Y_x + a'_3 \theta_2 \\ &= (a'_1 X + a'_2 x' + a'_3) \theta_2 + a'_1 \eta + a'_2 \eta_x + a'_1 \epsilon + a'_2 \epsilon_x.\end{aligned}$$

Now  $\epsilon$  and  $\epsilon_x$  are independent of each other and all the other terms, so any vector  $a$  with non-zero entries in either  $a_1$  or  $a_2$  must give rise to a random variable with positive variance. If these entries are all zero, but  $a$  is non-zero, then  $a$  must have at least one non-zero entry in  $a_3$ ; again, from the above equation, it is clear that the resulting random variable must have positive variance.

**Lemma 3.2:**  $(\sigma^2 I_n + V + XV_1 X')^{-1} = M^{-1} - M^{-1} X (X' M^{-1} X + V_1^{-1})^{-1} X' M^{-1}$ ,

where  $M = \sigma^2 I_n + V$ .

**Proof:** The desired inverse matrices exist because of the assumption that  $V$  and  $V_1$  are covariance matrices and hence are positive definite. The lemma then follows as a special case of the more general "matrix lemma" proved in Lindley and Smith [1972].

**Lemma 3.3:**  $V_1 X' (M + XV_1 X')^{-1} = (X' M^{-1} X + V_1^{-1})^{-1} X' M^{-1}$ .

**Proof:** By Lemma 3.2,

$$\begin{aligned}V_1 X' (M + XV_1 X')^{-1} &= V_1 X' [M^{-1} - M^{-1} X (X' M^{-1} X + V_1^{-1})^{-1} X' M^{-1}] \\ &= V_1 X' M^{-1} - V_1 X' M^{-1} X (X' M^{-1} X + V_1^{-1})^{-1} X' M^{-1} \\ &= V_1 X' M^{-1} - V_1 [X' M^{-1} X + V_1^{-1} - V_1^{-1}] (X' M^{-1} X + V_1^{-1})^{-1} X' M^{-1} \\ &= V_1 X' M^{-1} - V_1 X' M^{-1} + (X' M^{-1} X + V_1^{-1})^{-1} X' M^{-1} \\ &= (X' M^{-1} X + V_1^{-1})^{-1} X' M^{-1}.\end{aligned}$$

Theorem 3.2:

$$(3.6) \quad \lim_{\substack{\lambda \rightarrow \infty \\ \mathbf{V}_1^{-1} \rightarrow 0}} E\{Y_X/Y-Y\} = \mathbf{z}'(\mathbf{X}'\mathbf{M}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}^{-1}\mathbf{y} \\ + \mathbf{v}'[\mathbf{M}^{-1} - \mathbf{M}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{M}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}^{-1}]\mathbf{y} \text{ and}$$

$$(3.7) \quad \lim_{\substack{\lambda \rightarrow \infty \\ \mathbf{V}_1^{-1} \rightarrow 0}} E\{\theta_2/Y-Y\} = (\mathbf{X}'\mathbf{M}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}^{-1}\mathbf{y},$$

where  $\mathbf{M} = \sigma^2\mathbf{I}_n + \mathbf{V}$ .

Proof: These equations follow directly from (3.2) and (3.3) upon application of Lemma 3.3 and some simple algebra. The derivations are virtually identical, so we will prove only (3.7). From Theorem 3.1,

$$E\{\theta_2/Y-Y\} = \beta_0 + \mathbf{v}_1'\mathbf{X}'(\sigma^2\mathbf{I}_n + \mathbf{V} + \mathbf{X}\mathbf{V}_1\mathbf{X}')^{-1}(\mathbf{y} - \mathbf{X}\beta_0)$$

so by Lemma 3.3,

$$E\{\theta_2/Y-Y\} = \beta_0 + (\mathbf{X}'\mathbf{M}^{-1}\mathbf{X} + \mathbf{V}_1^{-1})^{-1}\mathbf{X}'\mathbf{M}^{-1}(\mathbf{y} - \mathbf{X}\beta_0)$$

and as  $\mathbf{V}_1^{-1}$  converges to 0, this clearly converges to

$$\beta_0 + (\mathbf{X}'\mathbf{M}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}^{-1}(\mathbf{y} - \mathbf{X}\beta_0) \\ = (\mathbf{X}'\mathbf{M}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}^{-1}\mathbf{y}.$$

Theorem 3.3: Denote  $\mathbf{V}^* = \tau\mathbf{R}^*$ . Then we have the following limiting forms as

$\lambda \rightarrow \infty$ :

$$(3.11) \quad \lim_{\lambda \rightarrow \infty} \lim_{\substack{\lambda \rightarrow \infty \\ \mathbf{V}_1^{-1} \rightarrow 0}} E\{Y_X/Y-Y\} = \mathbf{z}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

$$(3.12) \quad \lim_{\lambda \rightarrow \infty} \lim_{\substack{\lambda \rightarrow \infty \\ \mathbf{V}_1^{-1} \rightarrow 0}} E\{\theta_2/Y-Y\} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

If  $\mathbf{R}$  is non-singular, then we have the following limits as  $\lambda \rightarrow 0$ :

$$(3.13) \quad \lim_{\lambda \rightarrow 0} \lim_{\substack{\lambda \rightarrow 0 \\ \mathbf{V}_1^{-1} \rightarrow 0}} E\{Y_X/Y-Y\} = \mathbf{z}'(\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ + \mathbf{r}'\{\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}^{-1}\}\mathbf{y}$$

$$(3.14) \quad \lim_{\lambda \rightarrow 0} \lim_{\substack{\lambda \rightarrow 0 \\ \mathbf{V}_1^{-1} \rightarrow 0}} E\{\theta_2/Y-Y\} = (\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}^{-1}\mathbf{y}.$$

Proof: From Corollary 3.2.2,

$$\lim_{\lambda \rightarrow 0} E\{Y/X/Y=Y\} = z'(X'C^{-1}X)^{-1}X'C^{-1}y \\ + r'\{C^{-1} - C^{-1}X(X'C^{-1}X)^{-1}X'C^{-1}\}y$$

where  $C = \lambda I_n + R$

$$= z'(X'D^{-1}X)^{-1}X'D^{-1}y \\ + \lambda^{-1}r'\{D^{-1} - D^{-1}X(X'D^{-1}X)^{-1}X'D^{-1}\}y$$

where  $D = I_n + \lambda^{-1}R$ . As  $\lambda \rightarrow \infty$ ,  $D \rightarrow I_n$ , and (3.11) results. An identical derivation leads from (3.10) to (3.12). To obtain (3.13) and (3.14), note that if  $R$  is non-singular, then  $C^{-1} = (\lambda I_n + R)^{-1} \rightarrow R^{-1}$  when  $\lambda \rightarrow 0$ . Thus (3.13) and (3.14) follow directly from (3.9) and (3.10), respectively.

Theorem 3.4: Denote by  $\hat{Y}$  the predicted value vector. Then:

$$(3.15) \quad \hat{Y} = [\sigma^{-2}I_n + (V + XV_1X')^{-1}]^{-1}[\sigma^{-2}y + (V + XV_1X')^{-1}X\beta_0]$$

$$(3.16) \quad E\{\theta_2/Y=Y\} = [X'(\sigma^2I_n + V)^{-1}X + V_1^{-1}]^{-1}[X'(\sigma^2I_n + V)^{-1}y + V_1^{-1}\beta_0].$$

Proof: We will use Smith's model (2.1), along with the results of Lindley and Smith [1972], to prove this theorem. First, from (2.1a), we know that

$$E\{Y/\theta_1\} = \theta_1. \text{ Thus we can calculate } \hat{Y} \text{ as:}$$

$$\hat{Y} = E\{\theta_1/Y=Y\}.$$

Equation (3.15) is thus a special case of the theorem proved by Lindley and Smith (equations (12) and (13)). Similarly, we can use the lemma proved by Lindley and Smith to verify (3.16). Combining (2.1a) and (2.1b), we obtain the distribution of  $Y$  conditional on  $\theta_2$  without the mediating parameter

$\theta_1$ :

$$Y/\theta_2 \sim N(X\theta_2, \sigma^2I_n + V).$$

This equation, together with (2.1c), matches the assumptions of the lemma, and equations (7) and (8) of Lindley and Smith give (3.16).

Theorem 3.5:  $\hat{Y}$  solves the minimization problem: find  $u$  to minimize

$$(3.17) \quad (u-y)'(u-y) + (u-X\beta_0)'[\sigma^2(V + XV_1X')^{-1}](u-X\beta_0).$$

If  $V_1^{-1} = 0$ , (3.17) becomes:

$$(3.18) \quad (u-y)'(u-y) + \lambda u' [R^{-1} - R^{-1}X(X'R^{-1}X)^{-1}X'R^{-1}]u.$$

Moreover, the second term in (3.18) is 0 if and only if  $u \in \text{col}(X)$ .

Proof: As in the proof of Theorem 3.4, we use the fact that  $\hat{Y}$  is equal to the posterior expectation of  $\theta_1$ . Combining (2.1b) and (2.1c), we see that the prior distribution of  $\theta_1$  is:

$$\theta_1 \sim N(X\beta_0, V + XV_1X').$$

The conditional distribution of  $Y$  given  $\theta_1$  is given by (2.1a) as:

$$Y/\theta_1 \sim N(\theta_1, \sigma^2 I_n).$$

Applying Bayes' Theorem, we find that the posterior probability density function of  $\theta_1$  is proportional to:

$$\begin{aligned} & \exp\{-[\sigma^{-2}(Y-\theta_1)'(Y-\theta_1) + (\theta_1 - X\beta_0)'(V + XV_1X')^{-1}(\theta_1 - X\beta_0)]/2\} \\ & = \exp\{-Q(\theta_1)/2\}. \end{aligned}$$

This is clearly the p.d.f. for a normal distribution, and its expectation can be found by minimizing the quadratic form  $Q(\theta_1)$ . But minimizing  $Q(\theta_1)$  is equivalent to minimizing (3.17). Thus, the posterior expectation of  $\theta_1$ , and hence  $\hat{Y}$ , minimizes (3.17). To derive (3.18), we require the simple matrix identity  $(V + XV_1X')^{-1} = V^{-1} - V^{-1}X(X'V^{-1}X + V_1^{-1})^{-1}X'V^{-1}$ .

Substituting this into (3.17), we obtain:

$$\begin{aligned} & (u-y)'(u-y) + (u-X\beta_0)'[\sigma^2(V + XV_1X')^{-1}](u-X\beta_0) \\ & = (u-y)'(u-y) + (u-X\beta_0)'\sigma^2[V^{-1} - V^{-1}X(X'V^{-1}X + V_1^{-1})^{-1}X'V^{-1}](u-X\beta_0) \end{aligned}$$

and using the standardized form  $V = \sigma^2 R/\lambda$  and the assumption that  $V_1^{-1} = 0$ ,

$$= (u-y)'(u-y) + \lambda(u-X\beta_0)'[R^{-1} - R^{-1}X(X'R^{-1}X)^{-1}X'R^{-1}](u-X\beta_0).$$

But  $X'[R^{-1} - R^{-1}X(X'R^{-1}X)^{-1}X'R^{-1}] = [R^{-1} - R^{-1}X(X'R^{-1}X)^{-1}X'R^{-1}]X = 0$ , so that we obtain:

$$= (u-y)'(u-y) + \lambda u'[R^{-1} - R^{-1}X(X'R^{-1}X)^{-1}X'R^{-1}]u.$$

This is precisely (3.18). If  $u \in \text{col}(X)$ , then  $u = X\beta$  for some vector  $\beta$ .

Then the equations four lines above imply that the second term in (3.18) is 0. On the other hand, suppose that  $u$  is a vector such that the second term above is 0. Note that this term has the form  $u' Au$ , where  $A$  is a positive semi-definite matrix. The quadratic form can be 0 only if  $Au = 0$ . This then implies that:  $0 = RAu = [I_n - X(X'R^{-1}X)^{-1}X'R^{-1}]u$ . The matrix  $I_n - X(X'R^{-1}X)^{-1}X'R^{-1}$  is easily seen to be an idempotent matrix which projects into the orthogonal complement of  $\text{col}(X)$ . Thus, the above condition implies that  $u \in \text{col}(X)$ .

Theorem 4.1:

$$(4.1) \quad \text{Var}\{Y_X/Y-Y\} = \sigma^2 + v + z' [X'(\sigma^2 I_n + V)^{-1}X + V_1^{-1}]^{-1} z \\ - 2v'(\sigma^2 I_n + V + XV_1X')^{-1}XV_1z \\ - v'(\sigma^2 I_n + V + XV_1X')^{-1}v.$$

$$(4.2) \quad \text{Var}\{\theta_2/Y-Y\} = [X'(\sigma^2 I_n + V)^{-1}X + V_1^{-1}]^{-1}.$$

Proof: These equations, like those of Theorem 3.1, result from Lemma 3.1 by applying standard formulas for partitioned multivariate normal vectors.

Lemma 5.1: Given the prediction equation of Corollary 3.2.2, the residual vector is given by:

$$(5.5) \quad \hat{e}(\lambda) = B(\lambda)y,$$

where  $B(\lambda) = \lambda[C^{-1} - C^{-1}X(X'C^{-1}X)^{-1}X'C^{-1}]$  and  $C = \lambda I_n + R$ .

Proof: From (3.9), the predicted value vector, as a function of  $\lambda$ , is:

$$\hat{Y}(\lambda) = A(\lambda)y,$$

where  $A(\lambda) = X(X'C^{-1}X)^{-1}X'C^{-1} + R[C^{-1} - C^{-1}X(X'C^{-1}X)^{-1}X'C^{-1}]$ .

It is easy to verify that  $X(X'C^{-1}X)^{-1}X'C^{-1} = I_n - \lambda^{-1}CB(\lambda)$ . Substituting this identity into the equation above we obtain:

$$A(\lambda) = I_n - \lambda^{-1}CB(\lambda) + \lambda^{-1}RB(\lambda) \\ = I_n + \lambda^{-1}(R - C)B(\lambda).$$

But  $C = \lambda I_n + R$  so that:

$$A(\lambda) = I_n - B(\lambda).$$

The residual vector is thus given by:

$$\hat{e}(\lambda) = y - \hat{Y}(\lambda) = y - [I_n - B(\lambda)]y = B(\lambda)y.$$

**Theorem 5.1:** Define  $h(\lambda) = (y - X\beta_0)'(\sigma^2 I_n + \lambda^{-1}\sigma^2 R + X'V_1X')^{-1}(y - X\beta_0)$ , and let

$$h^*(\lambda) = \lim_{V_1^{-1} \rightarrow 0} h(\lambda). \quad \text{Then:}$$

$$(5.6) \quad h^*(\lambda) = [B(\lambda)y]'(I_n + \lambda^{-1}R)[B(\lambda)y]/\sigma^2, \quad \text{and}$$

$$(5.7) \quad \lim_{\lambda \rightarrow \infty} h^*(\lambda) = RSS/\sigma^2,$$

where RSS is the residual sum of squares from ordinary least squares regression.

**Proof:** We will apply Lemma 3.2, noting that  $M = \sigma^2 I_n + V = \lambda^{-1}\sigma^2 C$ . We then obtain:

$$\begin{aligned} h(\lambda) &= (y - X\beta_0)'(\sigma^2 I_n + \lambda^{-1}\sigma^2 R + X'V_1X')^{-1}(y - X\beta_0) \\ &= (y - X\beta_0)' \{ \lambda\sigma^{-2}C^{-1} - \lambda\sigma^{-2}C^{-1}X(X'\lambda\sigma^{-2}C^{-1}X + V_1^{-1})^{-1}X'\lambda\sigma^{-2}C^{-1} \}^* \\ &\quad (y - X\beta_0) \end{aligned}$$

and as  $V_1^{-1}$  converges to 0, this clearly converges to:

$$\begin{aligned} &\lambda\sigma^{-2}(y - X\beta_0)' \{ C^{-1} - C^{-1}X(X'C^{-1}X)^{-1}X'C^{-1} \} (y - X\beta_0) \\ &= y'B(\lambda)y/\sigma^2. \end{aligned}$$

Finally, it is easy to show that  $I_n + \lambda^{-1}R$  is a generalized inverse of  $B(\lambda)$ ; that is,  $B(\lambda)(I_n + \lambda^{-1}R)B(\lambda) = B(\lambda)$ . Substituting this into the last line and noting that  $B(\lambda)$  is symmetric gives us (5.6). In order to prove (5.7), we may calculate the limit of (5.6) directly. Alternatively, recall from Theorem 3.3 that as  $\lambda \rightarrow \infty$ , the Bayesian estimates converge to the ordinary least squares estimates. Correspondingly, the residual vector  $B(\lambda)y$  converges to the OLS residual vector. The weighting matrix clearly converges to the identity. Hence, (5.6) converges to the residual sum of squares for OLS regression.

Lemma 5.2: Given model (2.1), let  $\hat{\mathbf{Y}}$  denote the vector of predicted values.

$$\lim_{\lambda^{-1} \rightarrow 0} E\{(\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{I}_n + \lambda^{-1}\mathbf{R})(\mathbf{Y} - \hat{\mathbf{Y}})\} = (n-p)\sigma^2.$$

Proof: We use Theorem 3.1 to find  $\hat{\mathbf{Y}}$ , using the standardized notation

$$\mathbf{V} = \lambda^{-1}\sigma^2\mathbf{R};$$

$$\hat{\mathbf{Y}} = \mathbf{X}\beta_0 + (\lambda^{-1}\sigma^2\mathbf{R} + \mathbf{XV}_1\mathbf{X}')(\sigma^2\mathbf{I}_n + \lambda^{-1}\sigma^2\mathbf{R} + \mathbf{XV}_1\mathbf{X}')^{-1}(\mathbf{Y} - \mathbf{X}\beta_0).$$

Adding and subtracting  $\sigma^2\mathbf{I}_n$  from  $\lambda^{-1}\sigma^2\mathbf{R} + \mathbf{XV}_1\mathbf{X}'$ , we find:

$$\begin{aligned}\hat{\mathbf{Y}} &= \mathbf{X}\beta_0 + \mathbf{Y} - \mathbf{X}\beta_0 - \sigma^2(\sigma^2\mathbf{I}_n + \lambda^{-1}\sigma^2\mathbf{R} + \mathbf{XV}_1\mathbf{X}')^{-1}(\mathbf{Y} - \mathbf{X}\beta_0) \\ &= \mathbf{Y} - \sigma^2(\sigma^2\mathbf{I}_n + \lambda^{-1}\sigma^2\mathbf{R} + \mathbf{XV}_1\mathbf{X}')^{-1}(\mathbf{Y} - \mathbf{X}\beta_0), \text{ so that}\end{aligned}$$

$$\mathbf{Y} - \hat{\mathbf{Y}} = \sigma^2(\sigma^2\mathbf{I}_n + \lambda^{-1}\sigma^2\mathbf{R} + \mathbf{XV}_1\mathbf{X}')^{-1}(\mathbf{Y} - \mathbf{X}\beta_0), \text{ and}$$

$$\begin{aligned}E\{(\mathbf{Y} - \hat{\mathbf{Y}})(\mathbf{Y} - \hat{\mathbf{Y}})'\} &= \sigma^4(\sigma^2\mathbf{I}_n + \lambda^{-1}\sigma^2\mathbf{R} + \mathbf{XV}_1\mathbf{X}')^{-1}E\{(\mathbf{Y} - \mathbf{X}\beta_0)(\mathbf{Y} - \mathbf{X}\beta_0)'\} \\ &\quad (\sigma^2\mathbf{I}_n + \lambda^{-1}\sigma^2\mathbf{R} + \mathbf{XV}_1\mathbf{X}')^{-1} \\ &= \sigma^4(\sigma^2\mathbf{I}_n + \lambda^{-1}\sigma^2\mathbf{R} + \mathbf{XV}_1\mathbf{X}')^{-1},\end{aligned}$$

since from Lemma 3.1, we know that:

$$E\{(\mathbf{Y} - \mathbf{X}\beta_0)(\mathbf{Y} - \mathbf{X}\beta_0)'\} = (\sigma^2\mathbf{I}_n + \lambda^{-1}\sigma^2\mathbf{R} + \mathbf{XV}_1\mathbf{X}').$$

We then have that for any  $n \times n$  matrix  $\mathbf{W}$ :

$$\begin{aligned}E\{(\mathbf{Y} - \hat{\mathbf{Y}})' \mathbf{W} (\mathbf{Y} - \hat{\mathbf{Y}})\} &= \text{trace}[\mathbf{W} E\{(\mathbf{Y} - \hat{\mathbf{Y}})(\mathbf{Y} - \hat{\mathbf{Y}})'\}] \\ &= \sigma^4 \text{trace}[\mathbf{W}(\sigma^2\mathbf{I}_n + \lambda^{-1}\sigma^2\mathbf{R} + \mathbf{XV}_1\mathbf{X}')^{-1}].\end{aligned}$$

Applying Lemma 3.2 to the expression above, and using the fact that the limit

and the trace may be interchanged, it is easy to verify that:

$$\lim_{\lambda^{-1} \rightarrow 0} E\{(\mathbf{Y} - \hat{\mathbf{Y}})' \mathbf{W} (\mathbf{Y} - \hat{\mathbf{Y}})\} = \sigma^2 \text{trace}[\mathbf{W}\mathbf{B}(\lambda)].$$

To complete the proof we need only show that  $\text{trace}[(\mathbf{I}_n + \lambda^{-1}\mathbf{R})\mathbf{B}(\lambda)] = n-p$ .

But this follows immediately from the fact that  $(\mathbf{I}_n + \lambda^{-1}\mathbf{R})\mathbf{B}(\lambda)$  is an idempotent matrix which projects into the orthogonal complement of  $\text{col}(\mathbf{X})$ .

**Theorem 5.2:** Given the sampling model  $Y \sim N(X\beta, \sigma^2 I_n + \lambda^{-1} \sigma^2 R)$ , and assuming that  $\sigma^2$  is known, the appropriate lack of fit statistic for the model

$E\{Y\} = X\beta$  is:

$$(5.6) \quad h^*(\lambda) = [B(\lambda)Y]'(I_n + \lambda^{-1}R)[B(\lambda)Y]/\sigma^2.$$

The sampling theory distribution of (5.6) under this model is  $\chi^2_{n-p}$ .

**Proof:** Under the above sampling theory model, we use generalized least squares to find the predicted value vector:

$$\hat{Y} = X[X'(I_n + \lambda^{-1}R)^{-1}X]^{-1}X'(I_n + \lambda^{-1}R)^{-1}Y.$$

It is now easy to show that the corresponding residual vector can be written:

$$Y - \hat{Y} = (I_n + \lambda^{-1}R)B(\lambda)Y.$$

The appropriate lack of fit statistic for this model is the weighted residual sum of squares, with the weight matrix inversely proportional to the variance matrix:

$$\begin{aligned} (Y - \hat{Y})'(\sigma^2 I_n + \lambda^{-1} \sigma^2 R)^{-1}(Y - \hat{Y}) \\ &= [(I_n + \lambda^{-1}R)B(\lambda)Y]'(I_n + \lambda^{-1}R)^{-1}[(I_n + \lambda^{-1}R)B(\lambda)Y]/\sigma^2 \\ &= [B(\lambda)Y]'(I_n + \lambda^{-1}R)[B(\lambda)Y]/\sigma^2. \end{aligned}$$

**Corollary 5.2.1:**  $h^*(\lambda)$  is monotone increasing in  $\lambda$ .

**Proof:** Let  $u$  be any vector and consider the function:

$$f(\lambda) = u'(I_n + \lambda^{-1}R)^{-1}u.$$

Since  $R$  is a positive definite matrix, it is clear that for every  $u$ ,  $f(\lambda)$  is a monotone increasing function of  $\lambda$ . Let us denote by  $e(\lambda)$  the vector of residuals from using generalized least squares on the sampling theory problem of Theorem 5.2. We found in proving that theorem that

$$h(\lambda) = e(\lambda)'(I_n + \lambda^{-1}R)^{-1}e(\lambda)/\sigma^2. \text{ Now let } \lambda_2 > \lambda_1; \text{ we wish to show that}$$

$h(\lambda_2) > h(\lambda_1)$ . We have:

$$\begin{aligned} h(\lambda_2) &= e(\lambda_2)'(I_n + \lambda_2^{-1}R)^{-1}e(\lambda_2)/\sigma^2 \\ &> e(\lambda_2)'(I_n + \lambda_1^{-1}R)^{-1}e(\lambda_2)/\sigma^2 \end{aligned}$$

by the general monotonicity property described above for  $f(\lambda)$

$$> e(\lambda_1)'(I_n + \lambda_1^{-1}R)^{-1}e(\lambda_1)/\sigma^2 = h(\lambda_1)$$

because  $e(\lambda_1)$  is the residual vector for the generalized least squares estimator when  $\lambda = \lambda_1$ , and hence gives a smaller weighted residual sum of squares than does  $e(\lambda_2)$ .

#### References

- Anderson, T. (1958). An Introduction To Multivariate Statistical Analysis.  
New York: John Wiley and Sons, Inc.
- Aronszajn, N. (1950). Theory of reproducing kernels. Trans. Amer. Math. Soc., 68, 337-404.
- Blight, B. J. N. and Ott, L. (1975). A Bayesian approach to model inadequacy for polynomial regression. Biometrika, 62, 79-88.
- Box, G. E. P. (1954). The exploration and exploitation of response surfaces: Some general considerations and examples. Biometrics, 10, 16-60.
- Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modeling and robustness. Jour. Roy. Stat. Soc., A, 143, 383-430.
- Box, G. E. P. (1982). Choice of response surface design and alphabetic optimality. Yates Volume.
- Box, G. E. P. and Draper, N. R. (1959). A basis for the selection of a response surface design. Jour. Amer. Stat. Assoc., 622-653.
- Box, G. E. P. and Wilson, K. B. (1951). On the experimental attainment of optimum conditions. Jour. Roy. Stat. Soc., B, 13, 1-45.
- Box, G. E. P. and Youle, P. V. (1955). The exploration and exploitation of response surfaces: An example of the link between the fitted surface and the basic mechanism of the system. Biometrics, 11, 287-323.
- Chung, K. L. (1974). A Course In Probability Theory. New York: Academic Press, Inc.
- Draper, N. R. and Smith, H. (1981). Applied Regression Analysis, Second Edition. New York: John Wiley and Sons, Inc.
- Harrison, P. J. and Stevens, C. F. (1976). Bayesian forecasting. Jour. Roy. Stat. Soc., B, 38, 205-247.

- Kimeldorf, G. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. Jour. of Math. Anal. and Applic., 33, 82-95.
- Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model. Jour. Roy. Stat. Soc., B, 34, 1-41.
- Moler, C. (1981). MATLAB Users' Guide.
- Myers, R. H. (1976). Response Surface Methodology.
- O'Hagan, A. (1978). Curve fitting and optimal design for prediction. Jour. Roy. Stat. Soc., B, 40, 1-41.
- Plackett, R. L. (1950). Some theorems in least squares. Biometrika, 37, 149-157.
- Ryan, T. A., Joiner, B. L. and Ryan, B. F. (1981). MINITAB Reference Manual.
- Sacks, J. and Ylvisaker, D. (1978). Linear estimation for approximately linear models. Ann. Stat., 6, 1122-1137.
- Smith, A. F. M. (1973). Bayes estimates in one-way and two-way models. Biometrika, 60, 319-329.
- Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. Jour. Roy. Stat. Soc., B, 40, 364-372.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER #2474	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle)  BAYESIAN MODELS FOR RESPONSE SURFACES OF UNCERTAIN FUNCTIONAL FORM		5. TYPE OF REPORT & PERIOD COVERED Summary Report - no specific reporting period
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s)  David M. Steinberg		8. CONTRACT OR GRANT NUMBER(s)  DAAG29-80-C-0041
9. PERFORMING ORGANIZATION NAME AND ADDRESS Mathematics Research Center, University of 610 Walnut Street Wisconsin Madison, Wisconsin 53706		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Work Unit Number 4 - Statistics & Probability
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office P. O. Box 12211 Research Triangle Park, North Carolina 27709		12. REPORT DATE January 1983
		13. NUMBER OF PAGES 74
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report)  UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report)  Approved for public release; distribution unlimited.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number)  Response Surfaces; Bayesian Linear Models; Polynomial Regression; Spline Functions; Bias; Predictive Distribution; Lack of Fit Test; Kalman Filter.		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number)  Experimental response functions are often approximated by simple empirical functions such as polynomials. Several methods for modeling such responses which take into account this approximate nature are described and are shown to be essentially equivalent. The models all involve a Bayesian analysis which reflects prior experimental belief about the ability of the empirical approxi- mation to represent the true response function. The models are also related to Kalman filters. Implications of the models for statistical inference are		

DD FORM 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

(Abstract cont.)

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

ABSTRACT (continued)

examined with particular attention to estimating the response function. Numerical examples help illustrate the models. A general predictive check is developed to examine the consistency of the model with the observed data.

**DATE**  
**ILME**